2021-03

STUDENT ACHIEVEMENT INDICATORS AT DEFENSE LANGUAGE INSTITUTE FOREIGN LANGUAGE CENTER

Brenner, Ian A.

Monterey, CA; Naval Postgraduate School

NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

STUDENT ACHIEVEMENT INDICATORS AT DEFENSE LANGUAGE INSTITUTE FOREIGN LANGUAGE CENTER

by

Ian A. Brenner

March 2021

Thesis Advisor:	Samuel E. Buttrey
Co-Advisor:	Jonathan K. Alt
Second Reader:	Colby J. Smithmeyer

Approved for public release. Distribution is unlimited.

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			*Form Approved OMB No. 0704-0188*
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC 20503.			
1. AGENCY USE ONLY ***(Leave blank)***	**2. REPORT DATE** March 2021	**3. REPORT TYPE AND DATES COVERED** Master's thesis	
4. TITLE AND SUBTITLE STUDENT ACHIEVEMENT INDICATORS AT DEFENSE LANGUAGE INSTITUTE FOREIGN LANGUAGE CENTER			**5. FUNDING NUMBERS**
6. AUTHOR(S) Ian A. Brenner			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000			**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A			**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**
11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.			
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release. Distribution is unlimited.			**12b. DISTRIBUTION CODE** A

13. ABSTRACT (maximum 200 words)

The Defense Language Institute (DLI) trains most of the cryptologic language analysts (CLA) that perform translation and analysis of data to support the United States military and intelligence communities. Students take the Defense Language Proficiency Test (DLPT) when graduating, passing if they achieve a score of L2/R2 (2 on the Listening portion, 2 on the Reading). DLI has been ordered to improve its students' scores upon graduation. It seeks an improved model to screen applicants for the potential to achieve the new, more difficult grading benchmark of 2+ Listening, 2+ Reading. Former NPS student Jonathan Bermudez-Mendez looked into predicting student test scores based on grades, prior language experience, Defense Language Aptitude Battery (DLAB) scores, whether a student was recycled from a different language program, language category, and whether the student attended an immersion program, using stepwise logistic regression. We show that random forests and neural networks, especially the former, can improve on existing predictive models. We also investigate some univariate relationships based on prior language experience and show that many aspects of prior language exposure are statistically significantly related to the event of a student passing at the new benchmark.

14. SUBJECT TERMS DLI, Defense Language Institute, Defense Language Proficiency Test, DLPT, language, success, grades, Defense Language Aptitude Battery, DLAB, stepwise, logistic, regression, random forest, neural network, step-wise, model, categorical, numeric, data, ASVAB, Armed Services Vocational Aptitude Battery, classification table, goodness of fit, AUC, Area Under the Curve, sensitivity, specificity, ROC, Receiver Operating Characteristic			**15. NUMBER OF PAGES** 141
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UU

NSN 7540-01-280-5500 Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

STUDENT ACHIEVEMENT INDICATORS AT DEFENSE LANGUAGE INSTITUTE FOREIGN LANGUAGE CENTER

Ian A. Brenner
Lieutenant Commander, United States Navy
BS, United States Naval Academy, 2009

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
March 2021

Approved by: Samuel E. Buttrey
Advisor

Jonathan K. Alt
Co-Advisor

Colby J. Smithmeyer
Second Reader

W. Matthew Carlyle
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The Defense Language Institute (DLI) trains most of the cryptologic language analysts (CLA) that perform translation and analysis of data to support the United States military and intelligence communities. Students take the Defense Language Proficiency Test (DLPT) when graduating, passing if they achieve a score of L2/R2 (2 on the Listening portion, 2 on the Reading). DLI has been ordered to improve its students' scores upon graduation. It seeks an improved model to screen applicants for the potential to achieve the new, more difficult grading benchmark of 2+ Listening, 2+ Reading. Former NPS student Jonathan Bermudez-Mendez looked into predicting student test scores based on grades, prior language experience, Defense Language Aptitude Battery (DLAB) scores, whether a student was recycled from a different language program, language category, and whether the student attended an immersion program, using stepwise logistic regression. We show that random forests and neural networks, especially the former, can improve on existing predictive models. We also investigate some univariate relationships based on prior language experience and show that many aspects of prior language exposure are statistically significantly related to the event of a student passing at the new benchmark.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF FIGURES

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

ASVAB	Armed Services Vocational Aptitude Battery
AUC	Area Under the Curve
CAT	Language Category of Difficulty for languages taught at DLI
CLA	Cryptologic Language Analyst
CPH	Critical Period Hypothesis
DLAB	Defense Language Aptitude Battery
DLI	Defense Language Institute
DLIFLC	Defense Language Institute Foreign Language Center
DLPT	Defense Language Proficiency Test
DOD	Department of Defense
FY	Fiscal Year
GPA	Grade Point Average
ILR	Interagency Language Roundtable
LHS	Latin Hypercube Sampling
MOE	Measure of Effectiveness
NN	Neural Network
NPS	Naval Postgraduate School
NSA	National Security Agency
OLS	Ordinary Least Squares
OPI	Oral Proficiency Interview
RF	Random Forest
ROC	Receiver Operating Characteristics
SLA	Senior Language Authority
SDB	Student Database
SVM	Support Vector Machine
TRAC	The Research and Development Center
U.S.	United States of America
USCG	United States Coast Guard

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

Defense Language Institute Foreign Language Center (DLIFLC) is a Department of Defense (DOD) educational and research institute. It gives language instruction in 16 military-focused languages to 2,500 students each year (DLIFLC Catalogue 2020). At the end of instruction each student takes the Defense Language Proficiency Test (DLPT). The current passing standard on the DLPT is a score of L2/R2/S1+ (Defense Language and National Security Education Office 2020). Seventy percent of students met this standard in 2019 (Defense Language and National Security Education Office 2020). The Director of the National Security Agency has instituted a higher passing standard for DLIFLC in order to meet requirements for operational Cryptologic Language Analysts (CLA) (Department of the Army 2015). In order to meet the new standard incrementally, DLIFLC would like to improve its passing score to L2+/R2+/S1+ (DLIFLC 2017).

Our team developed two machine learning models, neural networks (NN) and random forests (RF), and compared them to a logistic regression model from previous research in effort to better predict the probability of a student's success while studying at DLIFLC. The model building also served to identify the modeling method which performs the best, to guide future study in this topic. We found random forest models to perform the best out of the three types of models.

Multivariate models are the best method for predicting student outcomes, but these models' inner workings are often difficult to understand. Univariate analysis is not as effective, but provides understandable differences in the predictors that can be used to shape policy. To this end, our team also investigated whether certain language backgrounds predisposed students to different levels of success in the languages studied at DLIFLC. Chi-squared tests were employed to answer questions like: Will a student with a Chinese or a Spanish background perform better on the Korean DLPT? We found that certain language backgrounds are indeed associated with better performance in the different languages taught at DLIFLC.

The chi-squared methods were extended to investigate the qualities of a student's prior language experience, namely the source where the student learned the language, the types of experience the student had using the foreign language before studying at DLIFLC, and the proficiency level the student had attained with that foreign language. Chi-squared methods were brought to bear on the importance of many of the other predictors present in the data set including student induction status, marital status, the use of DLAB waivers, motivation, the student's educational background, service branch, and years of service. These predictors were chosen for deeper study because these were used as primary predictors in at least one of the modeling methods. Our research shows that all these predictors are statistically significant in predicting the success of the students, some positively and others negatively.

References

Defense Language and National Security Education Office (2020) Department of Defense language codes list. Accessed January 13, 2021, https://dlnseo.org/sites/default/files/2020%20DoD%20Language%20Codes%20List%201-31-2020.pdf.

Defense Language Institute Foreign Language Center (2020) General catalog 2019–2020, v10c (Monterey, CA). https://www.dliflc.edu/wp-content/uploads/2018/11/DLIFLC_catalog_2019-20_v10c.pdf.

Defense Language Institute Foreign Language Center (2017) Institutional self-evaluation. Report, DLIFLC, Monterey, CA. https://www.dliflc.edu/wpcontent/uploads/2018/01/DLIFLC-Self-Study-December-2017_small.pdf.

Department of the Army (2015) Defense Language Institute Foreign Language Center (DLIFLC) plan to achieve 2+/2+ executive summary. Memorandum, Washington, DC.

ACKNOWLEDGMENTS

I need to put out a heartfelt thank you to my wife Kelly Brenner, supporting me through all those late nights working on my masters. It isn’t easy being two working adults and still striving for higher education and career progression.

I would like to convey my deep appreciation to Doctor Samuel Buttrey. His patient and experienced guidance facilitated the completion of this master’s thesis. Thank you for your assistance and understanding.

I give my thanks to Major Colby Smithmeyer and Doctor Jon Alt for facilitating the dialog between myself and Defense Language Institute and for keeping the project grounded. I would also like to thank the greater TRAC (The Research and Development Center) Monterey community for allowing this data set to be utilized and for its support through this process.

Lastly, I would like to extend my gratitude to the Gradate Writing Center for extremely patient reviews of all my written work. Their critical services have facilitated the graduation of untold many military officers that have been allowed to become too comfortable in military forms of writing...that rarely conforms to standards of English recognized by the academia or the greater outside world.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

This research provides an improved model for selecting student candidates for the military's foreign language training program at Defense Language Institute Foreign Language Center (DLIFLC). The primary goal is to improve the prediction of a student's success. This can provide a means for DLIFLC is to improve the likelihood of students achieving high scores on the Defense Language Proficiency Test (DLPT) upon completion of a student's course of instruction at DLIFLC (Defense Language Institute Foreign Language Center 2020). An additional benefit of this work is the ability to identify students who are struggling and to be able to divert additional resources to improve these student's overall outcomes on the DLPT. DLIFLC can utilize data-driven insights to shape policy and improve the quality of student graduates. Progress will happen through making data-informed changes to DLIFLC's institutional policies.

A. MISSION OF DLIFLC

The mission of DLIFLC is "to provide the highest quality culturally based foreign language education, training and evaluation to enhance the national security of the United States; and, as an Associate of Arts Degree and certificate granting Institution, DLIFLC is wholly committed to student service member success" (DLIFLC 2021).

DLIFLC teaches sixteen languages as of 2021 (DLIFLC 2020). These languages are divided into four categories (CAT) based on the empirical difficulty of learning that specific language. CAT IV languages are considered to be the most difficult and have the longest instructional period of 64 weeks. CAT III languages take 48 weeks to teach. CAT I and II languages are considered the easiest to learn and are taught in 36 weeks (DLIFLC 2020).

Students must maintain a "C" grade point average throughout the course of instruction. Students who are not able to meet this standard can be removed from the curriculum or may be restarted or relanguaged (starting the course of instruction in an easier language from the beginning.) There is also the additional option of a student being "recycled." This occurs when a student is returned to an earlier point in his or her respective language program, ideally to before when the student began having academic difficulty (DLIFLC 2017, p. 310).

At the end of a student's language instruction at DLIFLC, the student takes the DLPT and the Oral Proficiency Interview (OPI). The Department of Defense (DOD) standard for measuring language proficiency is the DLPT (Department of Defense 2009). Each portion of the exam is given an Interagency Language Roundtable (ILR) Language Skill Level. The ILR organizes language-associated activities at the federal level. There are six scoring levels on a scale of 0 to 5. Each score comes with a possible "+" modifier, for a total of 12 possible scores on each portion of the test. The three portions of the test consist of listening, reading, and speaking sections. The current graduation requirement from DLIFLC is to obtain a score of listening and reading skill level 2 (L2/R2) from the DLPT (DLIFLC Catalogue 2020).

B. THE PROBLEM

The DOD requires DLIFLC to elevate the DLPT standards for graduation to L2+/R2+/S1+ from L2/R2 by 2023 (Department of Defense 2009). The S1+ modifier is a score on the OPI. For the purposes of this research the L2+/R2+ (without S1+) standard is what was examined in the student data. The true requirement for CLAs working for the NSA is to improve student performance to L3/R3, but this is viewed to be beyond current reach. Unsuccessful students are very expensive to the DOD, both in dollar cost and the opportunity cost of that service member's time. DLIFLC can benefit from a deeper understanding of the factors that contribute to a student attaining a L2+/R2+ score for better screening of potential student candidates. We study secondary languages as indicators to see if specific language backgrounds produce a better aptitude for each of the DLIFLC's 16 offered languages. For example, a student with experience in Spanish might have better chances of achieving a L2+/R2+ on the Arabic DLPT than a student with experience in Hindi. Our team investigates the statistical performance of other predictors in the Student Database (SDB) set for achievement on the DLPT. A student's success is measured with a binary indicator showing whether he or she did, or did not, achieve at least a L2+/R2+ on the DLPT. The most recent analysis of factors to predict DLIFLC's student aptitude used stepwise logistic regression to generate the predictive model. Further in the past, decision trees and other machine learning methods have been used. We compare the use of random forests (RF) and neural networks (NN) to see if these models are able to more accurately predict student performance.

THIS PAGE INTENTIONALLY LEFT BLANK

II. BACKGROUND AND LITERATURE REVIEW

A. BACKGROUND

1. Mission of DLIFLC

DLIFLC provides foreign language instruction to train linguists for the federal government and military services. DLIFLC structures its instruction so that students will "understand and interpret meaning and intent with foreign language and culture including value systems, behavioral patterns, institutions, geography, and political, economic, and social systems of the areas where the target language is spoken" (Defense Language Institute Foreign Language Center 2019, p. 9). Its main mission is to produce linguists capable of doing Cryptologic Language Analyst (CLA) work at a proficiency level that is valuable to DOD (Defense Language Institute Foreign Language Center 2019).

The intelligence community identified a need for more proficient CLAs after the catastrophic terrorist attack of 9/11. Lieutenant General Michael Hayden, director of the National Security Agency (NSA), issued a memo in 2002 instructing all CLAs to score at least a L3/R3 on the DLPT (Department of the Army 2015). DLIFLC has attempted many actions to improve the proficiency levels of its linguistic graduates. DLIFLC has hired more instructors, adapted new instructional methods and materials, increased the student-faculty ratio from 5:1 to 3:1 for difficult languages and 4:1 for simpler languages, adopted team teaching, and utilized more immersion programs (Defense Language Institute Foreign Language Center 2019). The DOD Senior Language Authority (SLA) directed the DLIFLC to raise the graduation standards from L2/R2 to L2+/R2+ by the start of fiscal year (FY) 2023 in an effort to get to the NSA's L3/R3 skill requirement. The SLA felt the current L2/R2 standard left too wide of an operational gap to ultimately achieve the L3/R3 standard while at the CLA's first duty location, so the L2+/R2+ was implemented to narrow the gap. (Department of the Army 2015).

2. Defense Language Aptitude Battery (DLAB)

The Defense Language Aptitude Battery (DLAB) is an exam produced by the Army and used to screen all applicants to DLIFLC's language programs. The primary method of sorting students into the student's assigned languages is their individual respective score on the DLAB exam. A score of 95 is required for study in a Category I language (French or Spanish). A score of 100 is required to study a Category II language (Indonesian). A score of 105 is necessary for study in a Category III language (Hebrew, Persian Farsi, Russian, Tagalog, and Urdu). In order to study the most difficult, Category IV languages (Arabic-Modern Standard, Arabic-Egyptian, Arabic-Iraqi, Arabic-Levantine Syrian, Chinese-Mandarin, Japanese, Korean, and Pashto) a DLAB score of 110 is required (DLIFLC general catalog 2019–2020, 2020).

The highest possible score on the DLAB is currently 164 (DLABprep.com 2010). Students can be placed into the CAT one level above the one for which they qualified by waiver, but waivers are used only infrequently (DLABprep.com 2010). Waivers are usually reserved for applicants with some sort of prior language experience; one of DLIFLC's goals has been to limit the number of waivers in an effort to increase favorable DLPT outcomes. Each of the military service branches makes its own respective determination on who will take the DLAB based on an individual's performance on the ASVAB, education level, and testing availability (Schmitz et. al. 2009).

3. Armed Services Vocational Aptitude Battery (ASVAB)

The Armed Services Vocational Aptitude Battery (ASVAB) exam is administered to all prospective enlisted military members and many officers to measure the applicant's capacity for vocation and learning in four domains: verbal, math, science and technical, and spatial. The services adopted the ASVAB as the primary selection and classification tool in 1976. The ASVAB has undergone periodic changes with the last norming implemented in 2004 (Personnel Testing Division 2009).

In total the ASVAB consists of nine separate tests, depicted in Table 1. Scores on each test are used to determine eligibility for different vocations in the military. Most military vocations specifically screen for scores in certain areas and have an overall minimum passing score on the Armed Forces Qualification Test (AFQT). The two language-related tests on the exam are the Word Knowledge (WK) and the Paragraph Comprehension (PC) section. The WK measures the "ability to select the correct meaning of a word presented in context and to identify the best synonym for a given word" (ASVAB Fact Sheet 2020). The PC quantifies the applicant's "ability to obtain information from written passages" (ASVAB Fact Sheet 2021).

Table 1. Armed Services Vocational Aptitude Battery Subtest Descriptions. Source: Personnel Testing Division (2009)

Subtest	**Number of Items**	**Testing Time**	**Subtest Description**
General Science (GS)	25	11	Knowledge of physical and biological sciences
Arithmetic Reasoning (AR)	30	36	Ability to solve arithmetic word problems
Word Knowledge (WK)**	35	11	Ability to select the correct meaning of words presented in context and to identify the best synonym for a given word
Paragraph Comprehension (PC)**	15	13	Ability to obtain information from written passages
Math Knowledge (MK)	25	24	Knowledge of high school mathematics principles
Electronics Information (EI)	20	9	Knowledge of electricity and electronics
Auto & Shop Information (AS)	25	11	Knowledge of automobile technology, tools, and shop terminology and practices
Mechanical Comprehension (MC)	25	19	Knowledge of mechanical and physical principles
Assembling Objects (AO)*	25	15	Ability to determine how an object will look when its parts are put together
**Note: The Army does not use Assembling Objects (AO) to calculate composite test scores.*			
***Verbal Expression (VE) = weighted composite of WK and PC.*			

The AFQT is computed as a composite of three subtest scores and reported as a percentile. The services use the AFQT, to determine enlistment eligibility, (Personnel Testing Division, 2009):

$$AFQT \ = AR + MK + 2(VE)$$

The AFQT is used to determine admissibility for the four primary military services. The AFQT is reported as a percentile from 1 to 99. The current scores for the AFQT were last normalized in 1997 against a sample of 18–23 year-old youth that took the ASVAB (ASVAB Fact Sheet 2020). The AFQT is broken down into categories depicted in Table 2. An applicant who received a 90 on their AFQT would be in category II and scored as well as or better than 90% of his or her peers (ASVAB Scoring 2021). For an individual to become an Army CLA, he or she must score high enough on the AFQT to be eligible for military service, and score at least a 91 on the Skilled Technician composite test score. The Skilled Technician score is computed by adding the Verbal Expression, General Sciences, Mechanical Comprehension, and Math Knowledge portions of the ASVAB together (Personnel Testing Division 2009).

Table 2. AFQT Score Categories. Source: ASVAB Scoring (2021)

AFQT Category	Score Range
I	93-99
II	65-92
IIIA	50-64
IIIB	31-49
IVA	21-30
IVB	16-20
IVC	10-15
V	1-9

Cryptologic Linguists is the most common rating for students at DLIFLC. CLAs from all services must score a minimum of a 91 on the AFQT (Military.com 2020a). Navy Personnel must have VE + MK + GS score ≥ 162 and DLAB score ≥ 100 (ASVAB Scores and Navy Jobs 2020c). Air Force enlisted personnel must achieve a Verbal Expression (WK plus PC) and Arithmetic Reasoning (AR) score combined of at least 72 (Military.com 2020b).

DLIFLC Measure of Effectiveness

DLIFLC uses academic production as its foremost measure of effectiveness (MOE) to quantify improvement in DLPT scores. DLIFLC expresses academic production as:

$$\text{Academic production} = \frac{\text{Total \# of students achieving graduating standard}}{\text{Total enrolment-Administrative attrition}}.$$

The current L2+/R2+ production rate is approximately 36%. The goal for FY 2023 is to increase this to 64%. The rationale behind the 64% number is this: DLI would like 80% of students to pass the course of instruction (allowing for 10% academic attrition and 10% administrative attrition), and then to have 80% of the students passing the course also pass the DLPT with the L2+/R2+ requirement. If 80% of students pass the course and 80% of those students achieve the L2+/R2+ requirement, overall production will be 64%.

B. LITERATURE REVIEW

1. Previous Work

The Naval Postgraduate School (NPS) and DLIFLC have collaborated and produced numerous studies into academic performance and attrition of DLIFLC students. These studies built logistic regressions and decision tree models on the data with the processing ability that was available at the time of the model's inception.

2. Use of Tree Models for Predictive Modeling

Robert Anderson (1997) wrote a thesis titled "Study of Initial Student Attrition from Defense Language Institute Foreign Language Center." This thesis is important to our background research because it was the first and only documented use of decision trees to perform predictive analysis for the DLIFLC student data set. Our work builds on this, testing how successfully random forests, which are an ensemble of trees, perform when compared to other machine learning methods.

In his research Anderson utilized binary classification trees to build models with DLIFLC data from FY 1994 through FY 1996 in an attempt to model predictors of student attrition at DLIFLC. He did not document if his decision trees improved modeling of DLIFLC's data compared to other methods (Anderson 1997).

Anderson came to three primary conclusions through his modeling of the available data. The first primary conclusion was that a student's DLAB score was the best predictor of academic attrition for all languages studied except German. (German was a language taught at DLIFLC during this period of time, but it is no longer taught.) For students taking German the student's level of education was found to be the best predictor for likelihood to fail to complete DLIFLC's German language program, and DLAB the second best predictor (Anderson 1997).

He also found that the numerical DLAB score that produced the best split in the probability of passing or failure increased as the language categories increased (got harder). In most languages the student's service unit was the second best predictor for academic attrition. Army students were found to perform the worst. It should be noted that in some languages there were no Marine Corps or Navy students over several years. Generally the third most important predictor he found was the student's self-reported motivation for studying the language of instruction (Anderson 1997).

Anderson also found some predictive power in a student's ASVAB scores. The portions of the students' ASVAB scores that he had available were the AFQT, arithmetic reasoning, mathematics knowledge, and numerical operations subtests of the ASVAB. He did find that ASVAB scores for arithmetic reasoning were relatively good predictors for learning a foreign language compared to the other portions of the ASVAB scores he had available (Anderson 1997).

Anderson's second primary conclusion dealt with administrative attrition for students. He found the best predictor for administrative attrition to be the level of education for the students taking Russian and Arabic. He found a significantly different level of administrative attrition for students who had completed college or higher degrees than for those who that had not completed high school. He found that for students taking Spanish the administrative attrition was best modeled by the method of a student's entry to taking that language (Anderson 1997). Students who were regularly or administratively relanguaged students attrite at lower levels than students that were academically recycled.

The third primary conclusion was Air Force and the Navy administrators were more successful than the Marine Corps and Army in choosing which students should be restarted in the student's current language. The best predictors of administrative attrition for both regular and relanguaged inputs were subtests of the ASVAB (Anderson 1997).

3. Analysis of Factors Predicting Graduation of Students at DLIFLC

Chin Han Wong (2004) wrote a thesis titled "An Analysis of Factors Predicting Graduation of Students at Defense Language Institute Foreign Language Center" in 2004. He used a data set that spanned FY 1998 through FY 2003 and applied regression models for the analysis. Wong divided the students into different groups based on the language CAT that he or she were studying. He was able to show that non-relanguaged female Marines were the most likely successful graduates in all cases, and that Army males were the least likely to graduate. Students studying CAT III languages needed a 95 or greater on the DLAB to be likely to be successful and those studying CAT IV languages needed a score of 120 or greater.

Wong found interaction effects to be significant amongst the predictors of service, language CAT, gender, and DLAB scores. These findings indicate there are some complicated dynamics at play for the students in the different languages, and that members of some services handle the program better than others. Marine students had the highest graduation rates followed by members of the Navy, Air Force, and lastly Army (Wong 2004).

His research is interesting because many of these predictors were incorporated into the models built with random forests and neural networks in this paper. Models built using neural networks and random forests in many ways function like black boxes. It is difficult or impossible to understand what is happening to the data within the model's algorithms. Wong's research gives us some insight as to some of the interactions that are occurring. The research of this paper refutes Wong's claim that females outperform males. This discrepancy could be because our respective research utilizes very different DLIFLC datasets (Wong 2004).

For the students that were relanguaged, Air Force students did the best, followed by Army, Marines, and lastly Navy. Wong recommended that in future studies the Student Motivation and Retention Training (SMART) program be looked into as a predictor for success in future studies. SMART was a program that the services implemented around the turn of the millennium as a study skills and educational refresher program that some students were able to take (Wong, 2004).

4. Korean Academic Attrition at DLIFLC

Haupt (2004) studied the factors that caused attrition of students in the DLIFLC Korean language program in his thesis "Analysis of Korean Attrition at the Defense Language Institute Foreign Language Center." Korean is considered to be the most difficult language taught at DLIFLC. This study used the data from FY 2006 through FY 2013 to build regression models. His analysis showed the predictors of pay grade, service branch, DLAB scores, how a student was brought into the language program, prior language proficiency, and semester GPAs to be the best predictors of graduation success.

Haupt claimed that his models could predict whether a student will succeed or fail after the first semester with tolerable accuracy. The accuracy improved with data included from the second semester, but the third semester data did not appreciably improve the model's accuracy (Haupt 2014). The research conducted by Bermudez-Mendez seems to directly counter Haupt's claims (Bermudez-Mendez 2020). Bermudez-Mendez's first semester model was barely better than a coin flip in predicting student success. Additionally, the Bermudez-Mendez models did not gain significant accuracy until after the third semester's data was included. The work in this paper would seem to contradict Haupt as well.

5. Student Success Factors at DLIFLC

"Student Success Factors at Defense Language Institute Foreign Language Center" was a NPS thesis completed in March of 2020 by Jonathan Bermudez-Mendez. He identified the student factors that contributed to graduates achieving at least a 2+/2+ on the DLPT based on four separate student life cycle milestones: first day of classes, the end of the first semester, end of the second semester, and the end of the third semester. Student data for the study was taken from fiscal year (FY) 2008 through FY 2018 and fit to four separate stepwise logistic regression models.

The “day-one” model performed poorly at predicting success, but was very useful for identifying students who would not meet the new graduation standard. This model’s Receiver Operating Characteristic (ROC) area under the curve (AUC) was only 0.63. The models improved the predictive behavior as the semesters advanced towards graduation. By the third semester, Bermudez-Mendez was able to produce an AUC of 0.84, and the model was a strong predictor of a student’s success and failure. The models built in this paper will be directly compared to the Bermudez-Mendez third-semester logistic regression model (Bermudez-Mendez 2020).

The five important factors in predicting student success based on Bermudez-Mendez’s models were:

- Prior Language Experience: Students who already had experience as translators, transcribers, or language instructors had greatly improved odds of success in learning and showing proficiency on a new language in all models.

- Defense Language Aptitude Battery (DLAB): A student’s DLAB score was a significant predictor for a student’s performance on the DLPT throughout the student life cycle. Students who scored well on the DLPT had correspondingly high scores on the DLAB is most cases.

- Recycled Students: Recycled students were shown to have low odds of success on the DLPT.

- Language Category (CAT): Students in CAT 2, 3, and 4 languages had better odds of success than students in CAT 1 languages. It is possible that there are multiple reasons for this. For example: CAT 1 languages have shorter curriculum lengths and lower DLAB entry score requirements (Bermudez-Mendez 2020).

The student immersion program was not found to be a significant factor for predicting success on the DLPT. Students who were selected for these programs were already high-performing students with correspondingly high grade point averages (GPAs).

The models were controlled for GPA, and it was found that immersion was not a good predictor for success on the DLPT (Bermudez-Mendez 2020).

6. Age as a Predictor for Language Learning

Li F, Johnson J, and Yeung S (2014) published a paper entitled “Research on Age-related Factors in Foreign Language Learning” in the International Journal of Language and Linguistics. They were able to show that the primary predictor of whether a person could learn a second language to a level of proficiency comparable to the student’s native language was the person’s age. The younger the person is, the better he or she will be able to learn a second language. This falls in line with the traditional Critical Period Hypothesis (CPH) that humans learn language best before age nine, and the ability continues to drop significantly for most people past puberty (Li et al. 2014).

The CPH and associated research is interesting, because we found no evidence of age being a major predictor of student success at DLIFLC. There is no direct data category for age, but rank in the military is a pretty good proxy for age in the U.S. military. A student’s age can be deduced with good accuracy from the student’s rank. In all, rank proved to be a weak predictor in all models studied.

Shaw and Lett (1993) did a study entitled “Relationships of Language Aptitude and Age to DLPT Results among Senior Officer Students in DLIFLC Basic Language Courses.” In this study, DLIFLC students who had four or more years of prior military experience or more were compared to students who hado recently entered the service. When the language learning aptitude was controlled though the DLAB scores, it was found that length of service was not a significant a factor when compared to prior language experience (Shaw and Lett 1993). This is interesting because it would appear to counter the CPH. Shaw and Lett (1993) also found that prior language experience in Spanish was nearly twice as strong a predictor for DLPT success as was prior language experience with the more difficult languages of Russian, Arabic, and German (Shaw and Lett 1993).

This research is interesting because in the prior language studies conducted in this research students with Spanish backgrounds generally did not perform well in most of the languages studied at DLIFLC.

7. Predicting Aptitude for Foreign Language Attainment

Our team intended to study the feasibility of phasing out the DLAB exam in favor of utilizing the ASVAB exam in its place. If DLIFLC could use the ASVAB in place of the DLAB it would open up a much larger pool of potential students from which to choose. The statistical analysis for this line of questioning was completed and published by Lawrence A. Tomaziefski of Utica College as this team was gathering the data for study (Tomaziefski, 2020).

Tomaziefski (2020) found that the DLAB cannot easily be replaced by using ASVAB scores. He made a series of models to predict success on the DLPT based on ASVAB scores and various other demographic data that was available on each on each of the military trainee members. This study found a four-variable model using ASVAB Verbal Expression, Math Knowledge, Arithmetic Reasoning, and Auto and Shop Information scores to be more accurate in predicting DLPT scores than the models built on the ASVAB Skilled Technician score data. The models Tomaziefski built using ASVAB score and demographic data did not completely mitigate the risk of recruiting prospective CLAs with higher potential for attrition at DLIFLC without also using the DLAB.

This study into ASVAB scores recommended that the Army adopt the four-variable model as a screening tool at the initial point of entry and follow-up with DLAB testing either when the CLA candidate graduates or before entering foreign language training (Tomaziefski 2020). The purpose of this would be to better target students who would have potential for success on DLAB and on future success at DLIFLC as a whole. The DLAB exam is a relatively expensive program for the armed forces to produce and administer. There are relatively few testing sites that can handle proctoring the exam, so very few military personnel are able to even take the exam. With better targeting the armed forces might be able to raise the quality of applicants to DLIFLC (Tomaziefski 2020).

C. ANALYSIS METHODOLOGY

1. Random Forests

We utilized random forests to model the data. A random forest is a machine-learning model in which multiple decision trees are created and combined, making it a more powerful method for large, complex data sets. The combination is achieved through averaging to raise the predictive performance and reduce the variance in the ensemble (Breiman 2001).

Individual decision trees have the benefit of being an intuitive and easily understandable modeling technique. Decision tree's downside is that they tend to have high variance, over-fit the data, and often have weak predictive performance on test set data. A decision tree with few branches is thought to be a "weak learner": some predictive skill has occurred but the precision is low (Pham 2019).

Random forest models pool multiple decision trees, hopefully with low correlation among them, using bootstrap samples from the training data and other strategies. We used the software implementation of random forests for our investigations that is available in the R package (R core team 2020) "ranger" (Wright and Zeigler 2017).

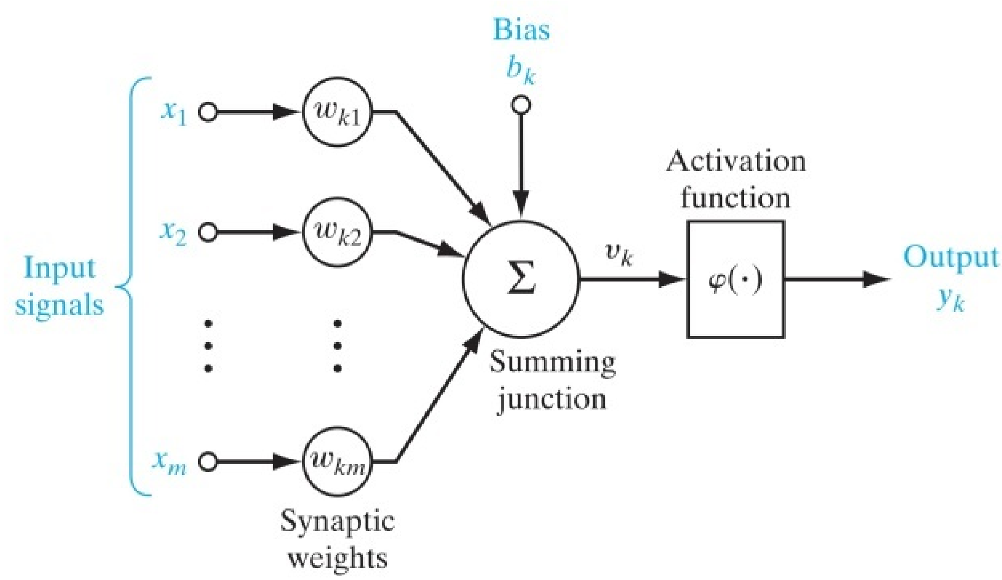
We include the best predictors and eliminate the least important ones to maximize the goodness of fit of our model. It is important to eliminate underperforming predictors so as not to produce an over-fit model. We use test-set prediction performance to guide us in selecting the right number of predictors to include. The goodness of a predictor for a categorical predicted value is usually measured based on improvement in the Gini index (Breiman et al. 1984) amassed over all bootstrapped trees for which the predictor is utilized in splitting. Using this criterion, we can rank the predictors from least to most important. We choose a subset of predictors that gave the best prediction performance on the test-set data (Pham 2019).

2. Neural Network

Artificial neural networks, usually called "neural networks," are an effort to mimic the organic neuron-based brain's ability to solve abstract problems (Goodfellow et al. 2016). The basic building blocks of neural networks are known as neurons. Neurons act as simple processing units for inputted information (predictors). Figure 1 depicts a basic neuron. The interconnection among these neurons creates a massively-parallel, distributed processor capable of learning features about the input predictors. The structure of a NN is based on a group of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain.

Neuron connections are referred to as "edges." The weights, the power of the interneuron connections, are used to store the learned information from the input data. The "weight" functions are numerical value the neuron attaches to the data it processes to signal the next layer's neuron. Thus, the output of one layer is the input of the next layer. The weight is computed as a non-linear function of the sum of its inputs. The weight raises or lowers the gain of the signal at a connection.

Figure 1. A Simplified Depiction of a Neuron in a Neural Network.
Source: Haykin (2009).



To produce a classification, a neuron requires the inputted values, the weights, and the activation function. The activation function, *f*, serves to make non-linear functions of the inputs. Bias is usually included as an additional parameter to adjust the net output for a fine-tuned result (Kim et. al. 2020). The output of a neuron can be mathematically shown as Figure 2:

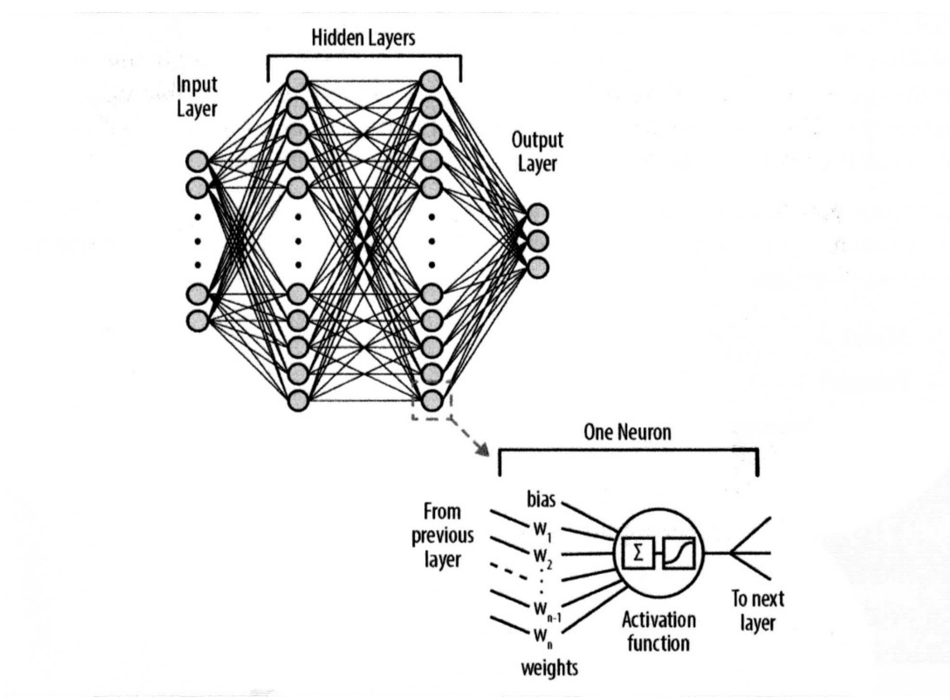
Figure 2. NN Neuron Output Function. Source: Kim et. al. (2020)

$$y_k = f(\sum_{j=1}^{m} W_{kj}X_j + b_k)$$

where y_k is the output of neuron k, m is the number of inputs, the X_j are the input signals; W_{kj} are the weights, and b_k is the bias (Haykin 2009).

Deep neural networks use several layers of neurons sandwiched between the initial input and final output, depicted in Figure 3. One of the steps of properly tuning a neural network is to adjust both the number of layers and number of neurons in each layer. The layer outputs become more refined before making the final classification as information passes through the layers (Goodfellow et al. 2016).

Figure 3. Network Layers, Neurons. Source: Cook (2017)



Normally neural networks are trained using stochastic gradient descent, which is an optimization method utilizing first-order derivatives to curtail the delta from comparing the neural network's final output to the true observations in the training set. This type of gradient descent is stochastic in nature in that it approximates the final gradient using subsets of the training data at each step. The gradients of each layer are found by recursive application of the chain rule in order to implement stochastic gradient descent on the multiple layers of a deep learning program (Li et al. 2020). Learning rates need to be chosen carefully. A learning rate that is too small will be more computationally intensive and take longer to converge into the final model. Learning rates that are too large are faster but may "overshoot" and diverge instead of converging (Bottou 2012).

We used the software implementation of neural networks for our research that is available in the R package (R core team 2020) "h2o" (LeDell 2020). H2O's Deep Learning is based on a multi-layer, acyclic artificial NN. The network is trained with stochastic gradient descent by adjusting the weights with respect to the error rate in the previous iteration. The network may have many hidden layers made up of neurons with tanh, rectifier, or maxout activation functions. Adaptive learning rate, rate annealing, momentum training, dropout, L1 or L2 regularization, check-pointing, and grid search are advanced features in h2o that can result in high predictive accuracy with careful use. Each node trains a copy of the global model parameters on its local data. That node will signal periodically to the global model with model averaging across the network (H2o.ai 2021).

3. Classification Tables

In each case of model building, only the first 75% of the data was used to build the models. This left the last 25% held in reserve for testing the models (logistic regression, random forest, and neural network). Each model's prediction output for the 25% test set was compared via classification tables and Receiver Operating Characteristic (ROC) curves to determine which model performed the best in predicting the students who achieved success on the DLPT at the 2+/2+ level.

Classification tables provide an easy way to visualize the predictive output of a classification model. These are constructed by laying out the predicted success and failures from the model against the true success and failures that occurred in the data set (Table 3). The differences between these values allow the analyst to calculate sensitivity, specificity, positive predictive value (PPV) aka precision, negative predictive value (NPV) and accuracy.

Sensitivity is a measure of how often we are accurate in predicting success for students that scored a 2+/2+ on the DLPT. Specificity is the measurement for correctly predicting failure for students who did not pass the DLPT with a 2+/2+ or above. Precision, or PPV, is the chances of a student truly succeeding if he or she is predicted to be successful in passing the DLPT at the 2+/2+ level or above. NPV is the chance of a student truly failing to pass at the 2+/2+ level if the student is predicted to fail by the model. Accuracy tells us how many times we accurately predicted success and failure in passing the DLPT at the 2+/2+ or above level (Sirsat 2021). The defining ratios for these values are included in Table 3.

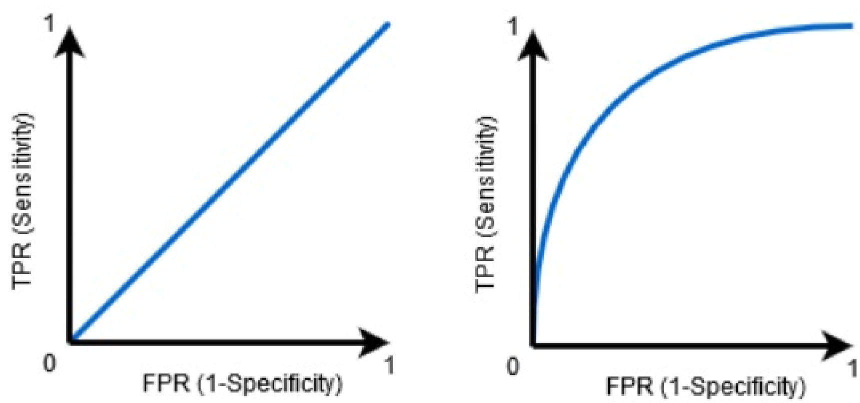
Table 3. Sample Classification Table. Source: Sirsat (2021)

		Predicted Class		
		Positive	**Negative**	
Actual Class	**Positive**	True Positive (TP)	False Negative (FN)	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP)	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

4. Receiver Operating Characteristic (ROC) Curves

Our team utilized ROC curves to assess the area under the curve (AUC) for our predicted values from the models. A ROC is plotted by plotting a model's sensitivity along the vertical axis of a two-dimensional plot and 1 – specificity, or the false positive rate (FPR), for the whole range of possible cutoff values. The AUC that is produced ranges from 0.5 to 1, and provides a visual metric of the model's ability to differentiate between students who pass at the 2+/2+ level on the DLPT and those who do not. Models with no predictive power will have AUC values close to 0.5. In this instance, random guessing is just as valuable. Models with an AUC of 1 can perfectly discriminate students who will be successful or fail at the DLPT (Bhandari 2020). Example ROC curves are shown in Figure 4.

Figure 4. Example ROC Curve. Source: Bhandari (2020)



The left ROC curve shows an AUC of .5, the right curve is above the 0.5 line.

5. Predictor Importance

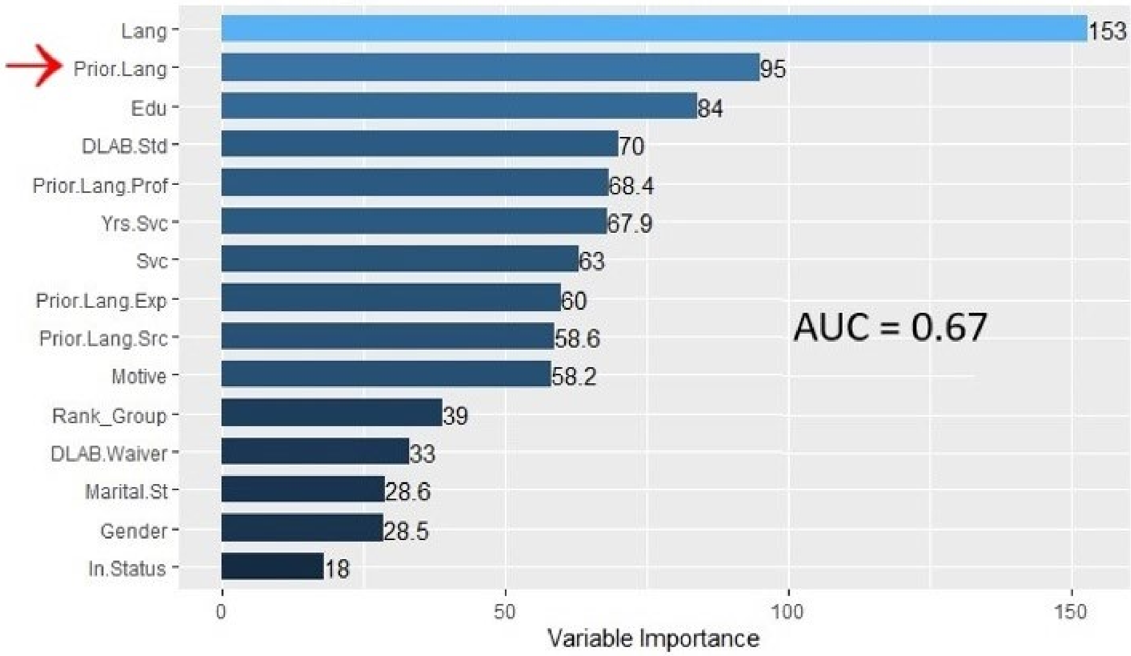
Initially the predictors used in all three models (logistic regression, random forest, and neural network) were the ones chosen by Bermudez-Mendez's third-semester logistic regression model (Bermudez-Mendez 2020). This was done to generate accurate comparisons between the three types of machine-generated models investigated in this body of work. The last step of comparison between the models will be determining the relative importance of the predictors chosen by the respective models when they are constrained to the ones in the Bermudez-Mendez model. Similarly, the ROC curves and truth table will also be compared for the un-constrained random forest and neural network.

6. Rationale for Prior Language Study

Our team made a random forest model containing only variables that can be known about a student when he or she first arrives at DLIFLC for training. The predicted variable indication made by this RF model is the student's associated probability of scoring L2+/R2+ or greater on the DLPT. RF is used to model this data, because RFs are shown to perform the best on this dataset later in this paper. This model produced an AUC=0.67. This is not a great AUC score. Only 39% of students pass at L2+/R2+, so this is not doing much better than guessing based off of the current pass rate. However the sensitivity is 0.90 (very high.) The model is useful for identifying those students who are likely to pass, and can highlight the most predictive variables. The most predictive "Day-0" variables are worth investigating for statistical significance and policy recommendations. The Day-0 RF model outperformed the Bermudez-Mendez logistic-regression (AUC=0.63) based on similarly chosen predictors (Bermudez-Mendez 2020). A graphic representing variable importance is shown in Figure 5.

Figure 5 shows that the prior language (Prior.Lang) of a student is the variable most associated with student success before a DLIFLC language of study (Lang) is chosen. The next most important associations in order are: the student's level of education on arrival, the student's standardized DLAB score, prior language proficiency, and the student's total years of service, branch of military service, the student's prior language experience, the student's prior language source, and the student's motivation for DLIFLC training. Variable importance tapers off after this point.

Figure 5. Day-0 RF Model Variable Importance



7. Language Groups

There exist 119 distinct language backgrounds for the students in the SDB data set. Most of the languages need to be grouped to develop a large enough sample size to be statistically useful. The language grouping was especially important to analyze the claim that a student's language background influences his or her retention in a given language of study. It became apparent later that combining the language backgrounds by grouping them into language groups made the models run much faster and improved the accuracy.

The languages were grouped along with others that have phonetic and grammatical structures. The grouping methodology was a mixture of research-driven language combinations and trial and error to see which languages behaved similarly across the data set. Many of the smaller groupings were found to be not useful and were combined into the generic "other" category to serve as a pseudo-control group as students with any non-English language background. This was especially the case with many of the languages found on the African continent which are very sparse in the data. The "other" category provides good contrast with the roughly one-quarter of the SDB data set comprised of students with no language background (other than English). Refer to Appendix D for a full listing of all the language digraph codes that went into the "other" language grouping.

Some languages like Spanish, Chinese, Arabic, and French were minimally grouped with other languages simply because the languages comprised a significant portion of the data set and had enough numbers for statistical analysis. Any additional grouping of these languages had the potential to diminish the results. Other languages like Hebrew, Japanese, and Korean were not grouped with any other languages because these languages are considered to be language isolates, completely unlike any other language within the data set (Augustyn 2018; Shibatani 2019; Martin 2019).

The bulk of the languages required much more work. The decision on how to group the Finnish language has been the most controversial of the decisions made in language grouping. Technically, Finnish is a language isolate belonging academically to the Finno-Ugric language group (Wordminds 2019). There have been many instances of linguistic drift in the Finnish language due to Europe's high level of trade, social mixing, and wars, and it appeared to behave like the other Germanic languages in testing. The decision was made to include this language in with what this research called the Germanic group. The Germanic group is made up of languages stretching across northern Europe such as German, Swedish, Dutch, Scottish, Norwegian, and Irish (Buccini 1998). The Balto-Slavic group includes languages such as Russian, Ukrainian, Slovak, Lithuanian, and Kashubian (Encyclopedia Britannica 1998).

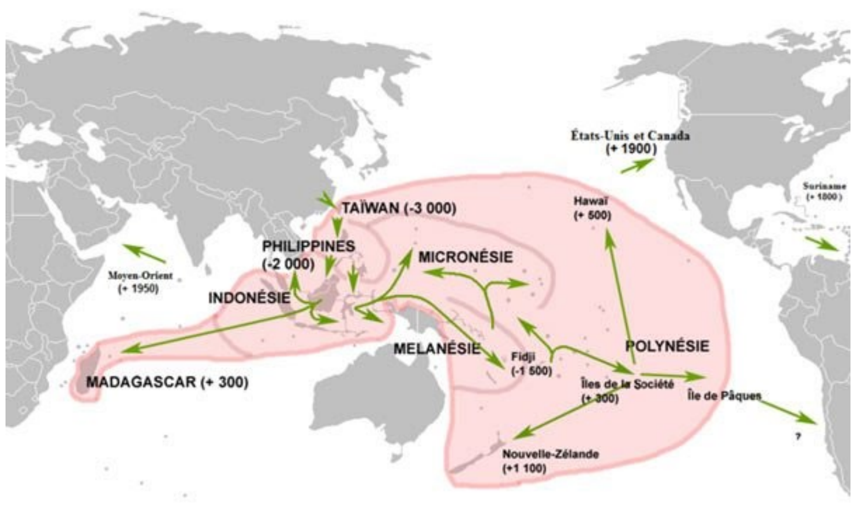
Although Spanish and French clearly fall within the Romance language group, these languages are excluded from being grouped in this way because Spanish and French are already large groups and have distinct behavior. The Romance group for this research includes the languages Italian, Catalan, Latin, Romanian, Albanian, Portuguese, and Provencal (Sala 1999).

The Indo-Iranian language group constitutes the languages that are the eastern-most of the Indo-European languages. The languages included in the Indo-Iranian group for this study are Persian-Iranian, Dari, Uzbek, Punjabi, Hindi, Oriya, Punjabi, Karanese, Gondi, Urdu, Pashto, and Kurdish (Cardona 1998).

Greek and Oghuz languages belong to academically separate language groups (Ruijgh et. al. 2018; Oghuz Languages 2021). These individual language groups are combined into one language group for this research due to the language's similar behavior in the data, and geographical proximity. The Greek-Oghuz group comprises the following languages: Greek, Turkish, Azerbaijani, Hungarian, and Armenian.

The Austroasiatic language group is a combination of the Cambodian, Thai, and Vietnamese languages (Diffloth 2018). The Austronesian language group is made up of Malagasy, Tagalong, Indonesian, Lao, Murano, Madurese, Palauan, Samoan, Malagasy, Kusaie, Waray-Waray, and Pampangan (Blust 2018). These relatively ancient languages originated around what is present-day Taiwan and spread through the Pacific, Indian, and even parts of the Caribbean as shown in Figure 6 (shown in the original French).

Figure 6. Austronesian Language Expansion. Source: Ancient Origins (2014)



8. Categorical Goodness of Fit Tests Using Chi-Squared Analysis

Our team decided to use chi-squared goodness of fit tests. These tests use two-way contingency tables to conduct the statistical analysis of many of the predictors in the SDB data set. This methodology was primarily employed to study the association of a student's prior language in contrast to the language studied and the overall proportion of students that were successful.

Table 4 shows an example contingency table employed in the language analysis. Here we have the L2+/R2+ pass and fail rates for students with a Korean background compared to the rates of students with only an English background. Contingency tables compare observed counts in each cell to the count which would be expected under the assumption of no relationship between row and column classifications (Devore 2016). The null hypothesis for a contingency table chi-squared test is that the proportions of pass and fail are the same in the population from which this sample is assumed to have been drawn. The null hypothesis is rejected when the two proportions observed in the students are quite different. The primary limiting assumption of this test is sample size: the usual rule of thumb is that there need to be at least five students expected in each cell of the table if the null hypothesis is true. The resulting p-value from this chi-square test in Table 4 is 1.259E-10, which is approximately zero. We reject the null hypothesis that students with Korean and English only backgrounds have the same passing rates on the DLPT. The mathematical representation of the chi-squared test and its underlying assumptions are displayed in Figure 7.

Table 4. Example Chi-Squared Contingency Table for Students with Korean Language Background across all languages taught at DLIFLC.

	Contingency Table	
chi-square p-value = ~ 0	**Pass**	**Fail**
Students with Korean Background	114	92
Students with an English-Only Background	1110	2239

Figure 7. Chi-Squared Goodness of Fit Test and Assumption. Source: Devore (2016)

Null hypothesis: $H_0 : p_1 = p_{10}, p_2 = p_{20}, \ldots, p_k = p_{k0}$
Alternative hypothesis: H_a : at least one p_i does not equal p_{i0}
Test statistic value: $\chi^2 = \sum_{\text{all cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \sum_{i=1}^{k} \frac{(n_i - np_{i0})^2}{np_{i0}}$
Provided that $np_{i0} \geq 5$ for all i, the P -value is (approximately) the area under the χ^2_{k-1} curve to the right of the calculated value of χ^2. If $np_{i0} < 5$ for at least one i, categories should be combined in a sensible way to correct this deficiency.

THIS PAGE INTENTIONALLY LEFT BLANK

III. DATA DESCRIPTION

A. DATA PREPARATION

The data preparation for this project went through three distinct phases. The first phase was cleaning the data for modeling. Approximately 70% of the work for this phase was accomplished by Bermudez-Mendez back in early 2020 before this present team was formed. The second phase was the grouping of the languages for conducting the language background comparisons. The third and final phase was analyzing the data and shaping it into the chi-squared statistical tables for statistical analysis.

In this chapter we will present the data variables that were used in either the modeling or the statistical analysis and the manipulations that had to happen to variables to make them usable. Many of the variables are unique to the DLIFLC and will be presented to afford the reader a level insight to the data set, and possibly help drive further study in this area.

DLIFLC's Directorate of Academic Administration keeps the Student Database (SDB). The data used for this analysis was exported from this SDB by Mr. Bryan Emerson of the Directorate of Academic Administration. This includes data from FY 2010 to FY 2018 and contains the DLIFLC historical student data for 26,714 students. The dataset has 53 different variables covering demographic and enrollment data for each student. Some of this data was generated by a questionnaire given to students when the student first arrives at DLIFLC. The questionnaire asks questions related to the student's prior educational and language experiences. The questionnaire data was recorded into the SDB if it was provided by the student. It should be noted that the majority of the data cleaning was done by Bermudez-Mendez. We built upon his work to produce this study.

The students that go on to serve as CLAs are the focus of this research. CLA students make up the majority of the student body at DLI and approximately 82% of the SDB data set. We removed all students from the data set who were academically or administratively separated early. These students never participated in the DLPT. Coast Guard and civilian students were also removed as not being the focus of the study. Students who studied in one of the languages no longer taught at DLIFLC were also removed from the data (Bermudez-Mendez 2020). A list of the 16 languages presently taught at DLIFLC can be found in Appendix C.

In some instances it was possible for students who did not complete the initial course of instruction to be relanguaged or otherwise retained in a different program. The records were filtered to look only at the final DLPT exam in the instances where there were multiple records for one student (Bermudez-Mendez 2020).

B. VARIABLE TRANSFORMATION

There were four types of variables that required deeper attention to prepare the SDB dataset for modeling. The first of these variables was the student rank in the military. This variable had 18 levels. It was grouped into three rank groups: junior enlisted, enlisted, and senior enlisted combined with officers. Approximately 75% of the data set is junior enlisted (E-1, E-2, and E-3). Enlisted is comprised of E-4, E-5, and E-6. Enlisted students make up close to 21% of the data set. Senior enlisted and officers make up the last 4% of the data set and span the ranks of E-7 through O-6 (Bermudez-Mendez 2020).

The FL series of variables give students' letter grade in one of the 15 classes taken DLIFLC. Grades range from A-F, and P's count as a pass in the subject. The grades themselves are 13 levels in total for the 15 FL series of variables. See Appendix B for the letter-grade-to-GPA conversion scheme that was used. After the grades were converted to numbers, the grades were combined in class groupings based on content. Classes that focused on foreign language skills were combined to form the FL1XX_Lang_Classes, FL2XX_Lang_Classes, and FL3XX_Lang_Classes. Similarly, the classes focused on culture and religion were combined to form the FL1XX_Culture_Classes, FL2XX_Culture_Classes, and FL3XX_Culture_Classes variables (Bermudez-Mendez 2020). The student's numerical score was a simple average of the classes that make up that specific group.

The prior language variable was the third type of variables that saw major transformations. This variable comes from the students' questionnaire form filled out when he or she starts DLIFLC. The student is asked what language other than English he or she has the most experience with. The student responses are coded into a two-digit, DOD recognized, alphabetic digraph code. This variable contained 119 levels before transformation. Thirteen of the levels were later found to be typographical errors and were combined into the "other language" category. The "other language" category was only supposed to be used for languages not yet recognized or included into the DOD's official list of recognized international languages (Defense Language and National Security Education Office 2020).

The fourth variable requiring transformation was the accumulated years of service of the service member. The data for this is strongly right-tailed, running from 0 to 26 years. The vast majority of students had a year or less of military experience when he or she reported to DLIFLC for training. There is a second peak in the data occurring right around year 6. For more information on this refer to Figure 17 and Table 15. When students with neighboring year groups behaved similarly, the years were grouped together. Outlier year groups were left ungrouped. The final grouping of year groups was: year 0, years 1–3, years 4–6, year 7, years 8–11, years 12–17, year 18, and years 19–26. For more information on this refer to Table 16.

Many languages were not recognized due to language's rarity or lack of use for the U.S. military. The remaining 98 prior language levels were combined into recognized "language groups" to make large enough groupings for statistical analysis. These language groups generally share similar grammatical and phonetic structures. In total all the languages were grouped into one of sixteen language groups.

For a full list of all the languages in the data set and how the languages were combined, refer to Appendix D. The reduced number of levels (16 versus 119) has the important benefit of making the computations in both the random forest and neural network significantly simpler and therefore faster in computational time, without losing much of the respective students' background.

It should be noted that in order to compute a neural network (NN) model all the variables had to be converted into discrete numerical values for the computation. Any rows of data with missing values in the predictors that were being utilized were left out for building the NN. This resulted in 6% less data for the NN.

C. PREDICTOR VARIABLES

Many of the predictor variables present in the data set were never used in the modeling process by any of the three models. These variables were largely comprised of dates and student serial numbers. Most of the variables used in this study are shown in Table 5. For the complete list of all variables contained in the dataset refer to Appendix A.

Table 5. Predictor Variables. Adapted from Bermudez-Mendez (2020)

Name	Symbol	Classification	Description
Service Branch	Svc	Categorical	USA (Army) USN (Navy) USMC (Marine Corps) USAF (Air Force)
Category	Lang.Cat	Categorical	Difficulty of Language: 1 (CAT I) 2 (CAT II) 3 (CAT III) 4 (CAT IV)
DLAB	DLAB	Continuous	Scores from 0–159
DLAB Waiver	DLAB_Waiver	Categorical	Y (yes) N (no)
Rank Group	Rank_Group	Categorical	Junior Enlisted (E-1, E-2, E-3) Enlisted (E-4, E-5, E-6) Sr. Enlisted & Officers (E-7 and above to O-6)
Years of Service	Yrs_Svc	Continuous	Originally years 0–26 This variable was transformed into the following year groups: year 0, years 1–3, years 4–6, year 7, years 8–11, years 12–17, year 18, and years 19–26

Name	Symbol	Classification	Description
Input Status	In_Status	Categorical	I (New Input) J (Relanguaged) P (Post-DLPT) Q (Recycle-Same Language)
Elementary Language Group	FL1XX_Lang_Classes	Categorical	Average Grade in FL101, FL102, and FL110
Intermediate Language Group	FL2XX_Lang_Classes	Categorical	Average grade in FL201, FL202, and FL210
Advanced Language Group	FL3XX_Lang_Classes	Categorical	Average grade in FL301, FL302, and FL310
Elementary Culture Group	FL1XX_Culture_Classes	Categorical	Average Grade in FL120 and FL140
Intermediate Culture Group	FL2XX_Culture_Classes	Categorical	Average grade in FL220 and FL240
Advanced Culture Group	FL3XX_Culture_Classes	Categorical	Average grade in FL320 and FL3240
Language Studied	Lang	Categorical	PV (Pashtu-Afghan) CM (Chinese-Mandarin) JA (Japanese) DG (Arabic-Iraqi) AE (Arabic-Egyptian) AP (Arabic Levantine-Syrian) KP (Korean) AD (Arabic Modern Standard) RU (Russian) UR (Urdu) TA (Tagalong) HE (Hebrew) PF (Persian-Farsi) JN (Indonesian) QB (Spanish) FR (French)
Student's Gender	Gender	Categorical	M (male) F (female)
Student's Motivation	Motive	Categorical	1 (Not motivated, does not want to study any language)

Name	Symbol	Classification	Description
			2 (Not motivated, prefers another language) 3 (Not my preferred language, but motivated to learn) 4 (Motivated, language is second or third choice) 5 (Motivated, language is first choice)
Marital Status	Marital.St	Categorical	S (single) M (married)
Prior Language	Prior.Lang	Categorical	16 language groups, originally 119 languages,
Language Immersion	Immersion	Categorical	O (OCONUS Immersion) C (CONUS Immersion) U (Immersion location not provided) *Blank* (No Immersion)

D. RESPONSE VARIABLE

The categorical variable of a student passing the DLPT with a score of L2+/R2+ or greater is the response variable for the study. This column of data was added to the original provided data for our use in model building. The S1+ OPI modifier was not included in this study. See Table 6 for a description of the response variable.

Table 6. Response Variable. Source: Bermudez-Mendez (2020)

Name	Symbol	Classification	Description
L2+/R2+ or Greater	L2+/R2+_greater	Categorical	0 (Did not achieve L2+/R2+) 1 (Achieved L2+/R2+ or greater)

E. DESCRIPTIVE STATISTICS

Descriptive statistics are presented in this section so that the reader can have an improved understanding of the distributions in the data. This will also afford the reader insight into what the statistical analysis presented in Chapter 4 tells us.

1. Students with Multiple Observations in the Dataset

134 students appear twice in the SDB data set (for a total of 268 observations). These students attended DLIFLC on two different occasions, taking different languages, and several years apart. The remaining 14,844 observations are individual students (Bermudez-Mendez 2020).

2. Distribution of Students Achieving L2+/R2+ or Greater

A total of 9,096 (61%) of the 14,896 observations in the SDB dataset did not achieve a score of L2+/R2+ or greater on the DLPT exam. In contrast 5,800 students (39%) did.

3. Student Distribution by Branch of Military Service

Figure 8 shows the numbers of students from each military branch. Table 7 displays the distribution table for the branches of service. Air Force students are by far the largest group with nearly half of all students. Army and Navy students have similar numbers, and Marine Corps students make up the smallest group. In Table 7 and the similar ones that follow we show the count of students by Pass (at the L2+/R2+ standard), Fail, and the Total. The Prop column shows the proportions passing in each row, and the PropPass shows the proportion of the total passing students contained in each row. It should also be noted that the Total displayed in some of the following contingency tables will be different depending on the data cleaning required for that specific predictor.

Figure 8. Student Distribution by Branch of Military Service

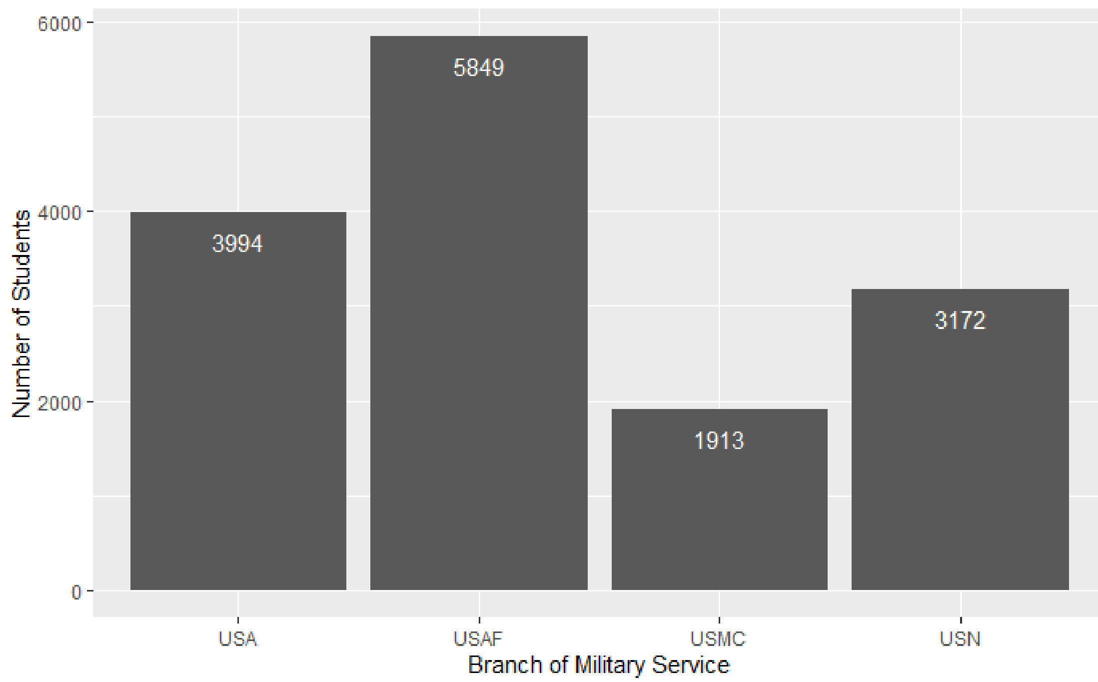


Table 7. Contingency Table for Student Branch of Military Service

Service	Total	Pass	Fail	Prop	PassRate	PropPass
USA	3994	1403	2591	26.80%	35.10%	24.10%
USAF	5849	2401	3448	39.20%	41.00%	41.30%
USMC	1913	752	1161	12.80%	39.30%	12.90%
USN	3172	1256	1916	21.20%	39.60%	21.60%
Overall	14928	5812	9116	100.00%	38.90%	100.00%

4. Student Distribution by Language Category

Figure 9 shows the distribution of students by language category. The Category IV languages contain the most students, at 54% of the total student population; Cat III languages are next with 28%; CAT I has 15%; and lastly CAT II has 1% of the student population. The distribution by language category of the students that pass the DLPT at the 2L+/2R+ level is depicted in Table 8.

Figure 9. Student Distribution by Language Category

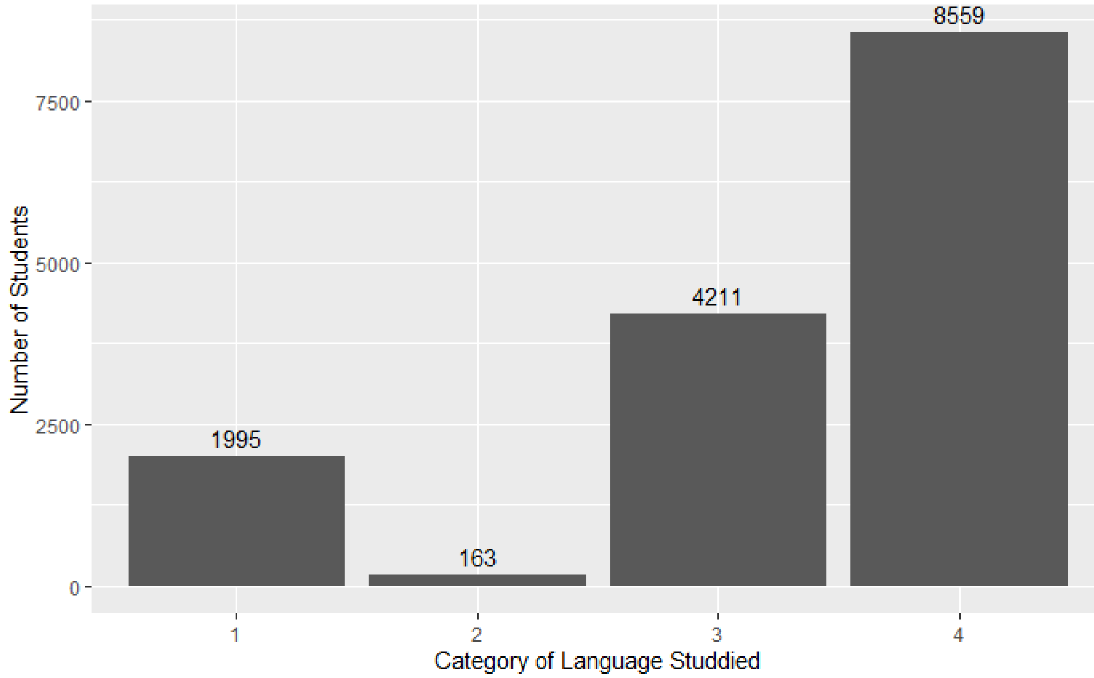


Table 8. Contingency Table for Language Category

Language Category	Total	Pass	Fail	Prop	PassRate	PropPass
1	1995	585	1410	13.36%	29.30%	10.10%
2	163	122	41	1.09%	74.80%	2.10%
3	4211	1595	2616	28.21%	37.90%	27.40%
4	8559	3510	5049	57.34%	41.00%	60.40%
Overall	14928	5812	9116	100.00%	38.90%	100.00%

5. Student Distribution by Standardized DLAB Score

The primary method of sorting students into a language category of difficulty is the DLAB score, not by any other informed metric. The data showed that all students with the same standardized DLAB score were in the same language category; the only exceptions were students with DLAB waivers.

6. Student Distribution by Prior Language Group

Figure 10 shows the student distribution by prior language grouping (including the None-English language group). Table 9 is the contingency table for prior language group.

Figure 10. Student Distribution by Prior Language Group

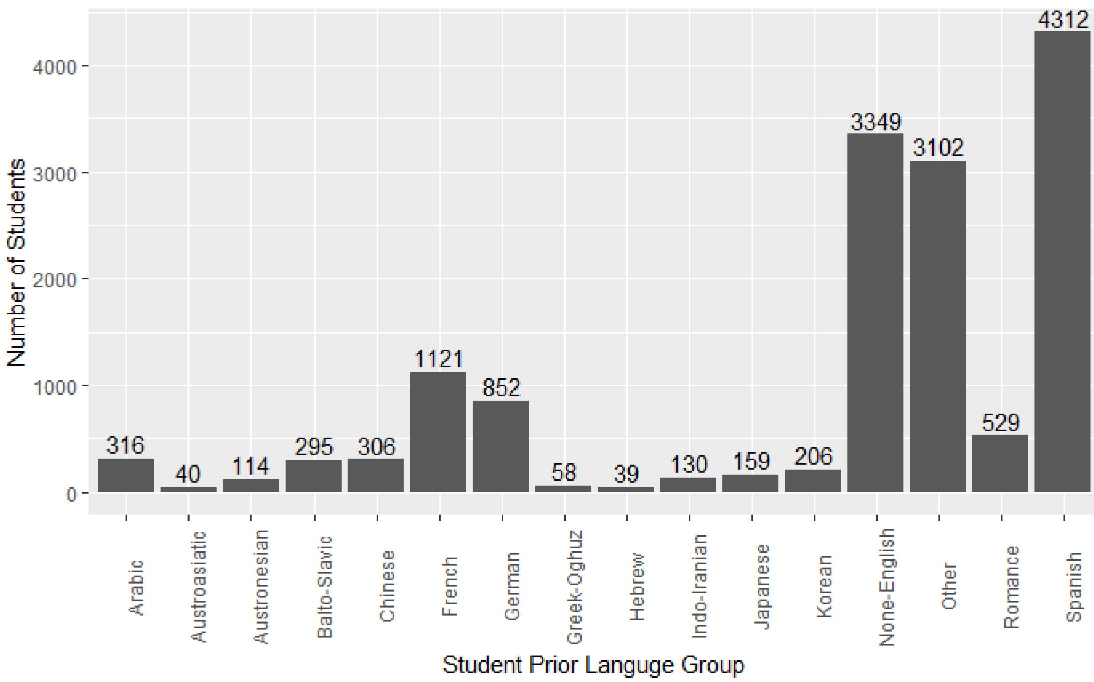


Table 9. Contingency Table for Prior Language Group

Prior Language Group	Total	Pass	Fail	Prop	PassRate	PropPass
Austroasiatic	40	20	20	0.27%	50.00%	0.34%
Arabic	316	166	150	2.12%	52.50%	2.86%
Austronesian	114	59	55	0.76%	51.80%	1.02%
Balto-Slavic	295	125	170	1.98%	42.40%	2.15%
Chinese	306	162	144	2.05%	52.90%	2.79%
French	1121	483	638	7.51%	43.10%	8.31%
German	852	365	487	5.71%	42.80%	6.28%
Greek-Oghuz	58	31	27	0.39%	53.40%	0.53%
Hebrew	39	21	18	0.26%	53.80%	0.36%
Japanese	159	65	94	1.07%	40.90%	1.12%
Korean	206	114	92	1.38%	55.30%	1.96%
Spanish	4312	1660	2652	28.89%	38.50%	28.56%
Indo-Iranian	130	54	76	0.87%	41.50%	0.93%
Romance	529	226	303	3.54%	42.70%	3.89%
Other	3102	1150	1952	20.78%	37.10%	19.79%
None-English	3349	1111	2238	22.43%	33.20%	19.12%
Overall	14928	5812	9116	100.00%	38.90%	100.00%

7. Student Distribution by Prior Language Source

Figure 11 shows the SDB data student distribution by prior language source prior to reporting for study at DLIFLC (including the students that reported “None”). Table 10 shows the contingency table for the student prior language source. It should be noted that the students were left to interpret and answer the questionnaire’s questions as they saw fit. This is the best rational our team understands to explain why the response numbers for “None” do not match throughout the different facets of the student prior language background. For instance, if a student learned a language in a work environment, but did not see that as one of the keyed for responses, they may have marked “None.”

Figure 11. Student Distribution by the Source of Prior Language Source

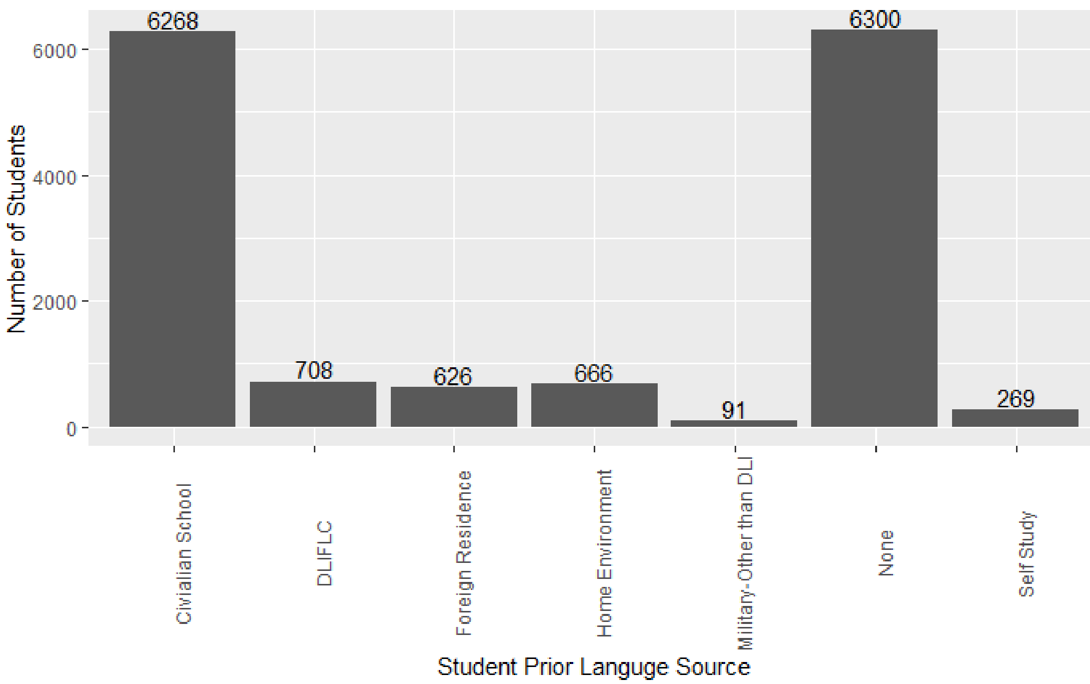


Table 10. Contingency Table for Source of Prior Language Source

Student Prior Language Source	**Total**	**Pass**	**Fail**	**Prop**	**PassRate**	**PropPass**
Civilian School	6268	2493	3775	41.99%	39.80%	42.89%
DLIFLC	708	334	374	4.74%	47.20%	5.75%
Foreign Residence	626	304	322	4.19%	48.60%	5.23%
Home Environment	666	317	349	4.46%	47.60%	5.45%
Military Schooling other than DLIFLC	91	50	41	0.61%	54.90%	0.86%
Self-Study	269	116	153	1.80%	43.10%	2.00%
None	6300	2198	4102	42.20%	34.90%	37.82%
Overall	14928	5812	9116	100.00%	38.90%	100.00%

8. Student Distribution by Prior Language Proficiency

Figure 12 depicts the student distribution by self-reported prior language proficiency upon checking into DLIFLC for study (including the "None" category). Table 11 shows the passing rates for students based on proficiency in prior foreign language.

Figure 12. Student Distribution by Proficiency in Prior Foreign Language

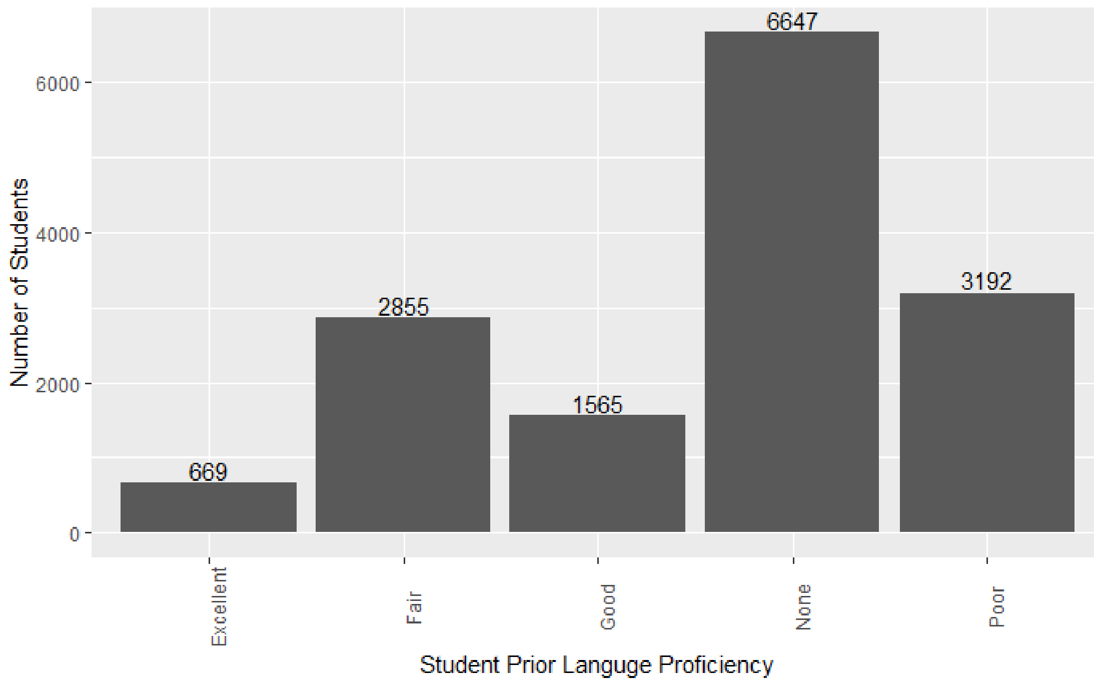


Table 11. Contingency Table Proficiency in Prior Foreign Language

Student Prior Language Proficiency	Total	Pass	Fail	Prop	PassRate	PropPass
Poor	3192	1222	1970	21.38%	38.30%	21.03%
Fair	2855	1182	1673	19.13%	41.40%	20.34%
Good	1565	738	827	10.48%	47.20%	12.70%
Excellent	669	355	314	4.48%	53.10%	6.11%
None	6647	2315	4332	44.53%	34.80%	39.83%
Overall	14928	5812	9116	100.00%	38.90%	100.00%

9. Student Distribution by Prior Language Experience in Usage

Figure 13 depicts the student distribution of the SDB data for the student's self-reported experience using a previously learned foreign language upon arriving for training. Table 12 depicts the contingency table for student prior language experience.

Figure 13. Student Distribution by Experience Using Prior Foreign Language

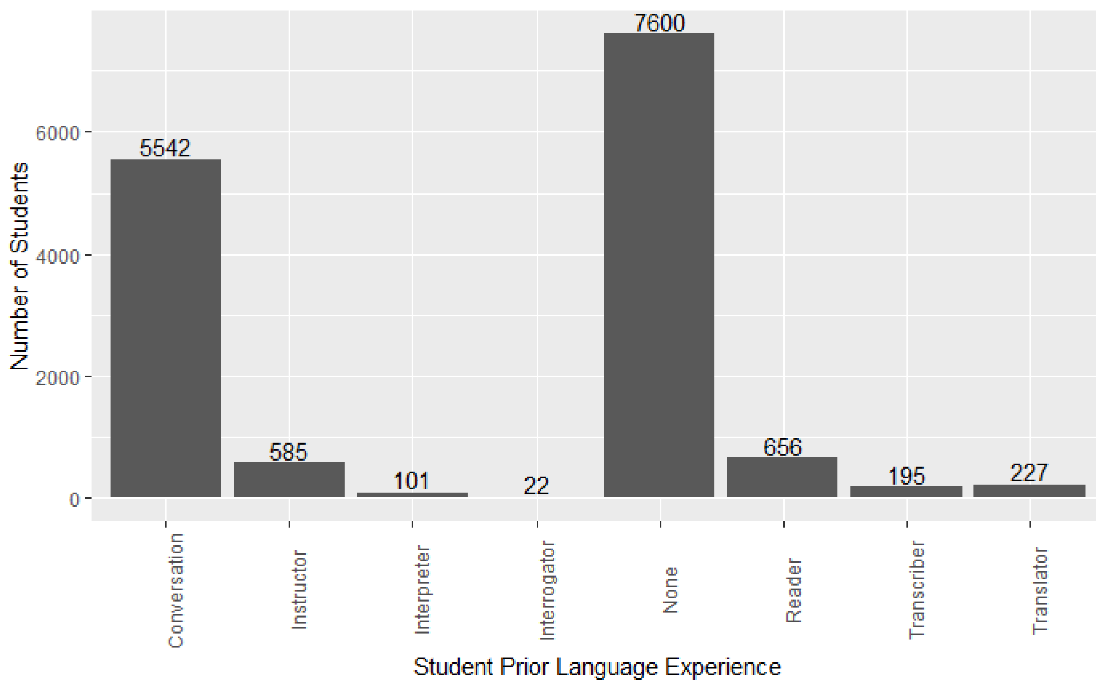


Table 12. Contingency Table for Student Experience Using Prior Foreign Language

Student Prior Language Experience	Total	Pass	Fail	Prop	PassRate	PropPass
Conversation	5542	2300	3242	37.13%	41.50%	39.57%
Instructor	585	270	315	3.92%	46.20%	4.65%
Interpreter	101	41	60	0.68%	40.60%	0.71%
Interrogator	22	8	14	0.15%	36.40%	0.14%
Reader	656	295	361	4.39%	45.00%	5.08%
Transcriber	195	106	89	1.31%	54.40%	1.82%

Student Prior Language Experience	Total	Pass	Fail	Prop	PassRate	PropPass
Translator	227	134	93	1.52%	59.00%	2.31%
None	7600	2658	4942	50.91%	35.00%	45.73%
Overall	14928	5812	9116	100.00%	38.90%	100.00%

10. Student Distribution by Prior Level of Education

Figure 14 depicts the SDB student distribution by the student's prior level of education. Table 13 depicts the contingency table for the same distribution of students.

Figure 14. Student Distribution by Prior Level of Education

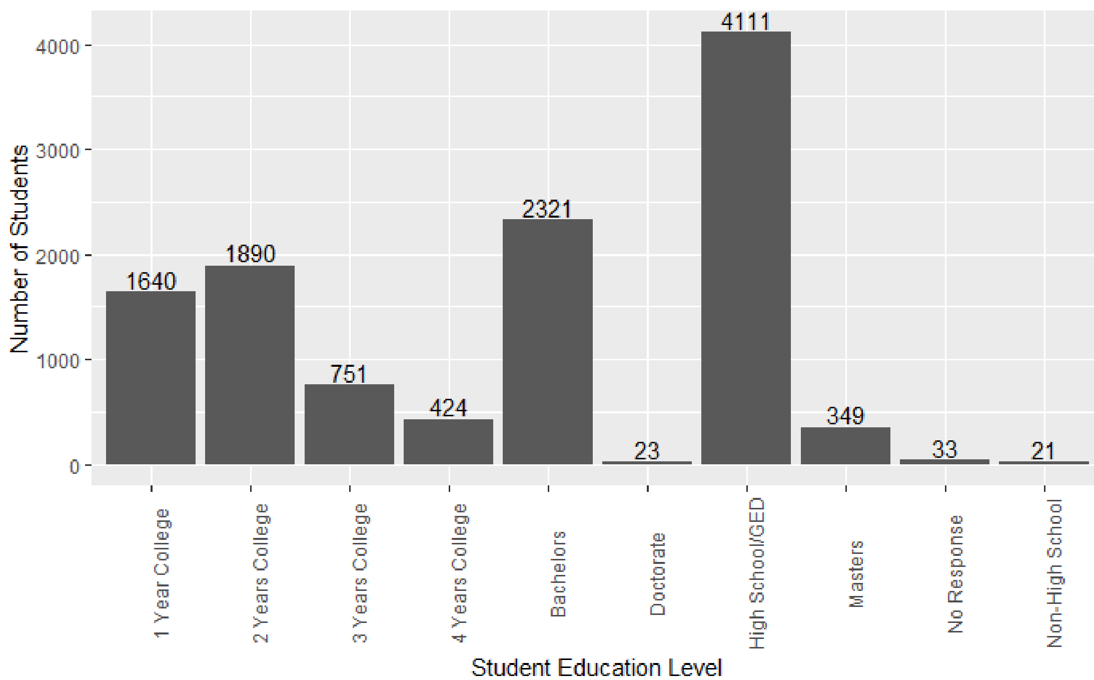


Table 13. Contingency Table for Student Prior Level of Education

Student Education Level	Total	Pass	Fail	Prop	PassRate	PropPass
No Response	33	9	24	0.29%	27.30%	0.19%
Non-High School	21	10	11	0.18%	47.60%	0.21%
High School/GED	4111	1521	2590	35.55%	37.00%	32.38%
1 Year College	1640	650	990	14.18%	39.60%	13.84%
2 Years College	1890	802	1088	16.35%	42.40%	17.08%
3 Years College	751	299	452	6.50%	39.80%	6.37%
4 Years College	424	184	240	3.67%	43.40%	3.92%
Bachelors	2321	1053	1268	20.07%	45.40%	22.42%
Masters	349	162	187	3.02%	46.40%	3.45%
Doctorate	23	7	16	0.20%	30.40%	0.15%
Overall	11563	4697	6866	100.00%	40.60%	100.00%

11. Student Distribution by Immersion Experience

Figure 15 depicts the distribution of students in the SDB data set by each student's access to DLIFLC sponsored immersion programs. The "C" column represents students who attended an immersion program that is within the continental U.S. The "N" column depicts students that did not participate in an immersion program. The "O" column depicts students that attended an immersion program that was outside the continental U.S. The "U" column shows students that attended an immersion program, but for whom there was not enough recorded data to determine what type.

Table 14 displays the contingency table for the same distribution as Figure 15. It would initially appear that students with immersion programs perform better than students without immersion programs. Bermudez-Mendez (2020) showed in his research that once GPA is controlled for, the students with immersion are not any more likely to be successful on the DLPT. Only the top academic students are chosen for the costly immersion programs, and many of them would likely still have scored well on the DLPT with or without the immersion program. No data exists for what would happen if the immersion program was used on struggling students.

Figure 15. Student Distribution by Immersion Experience

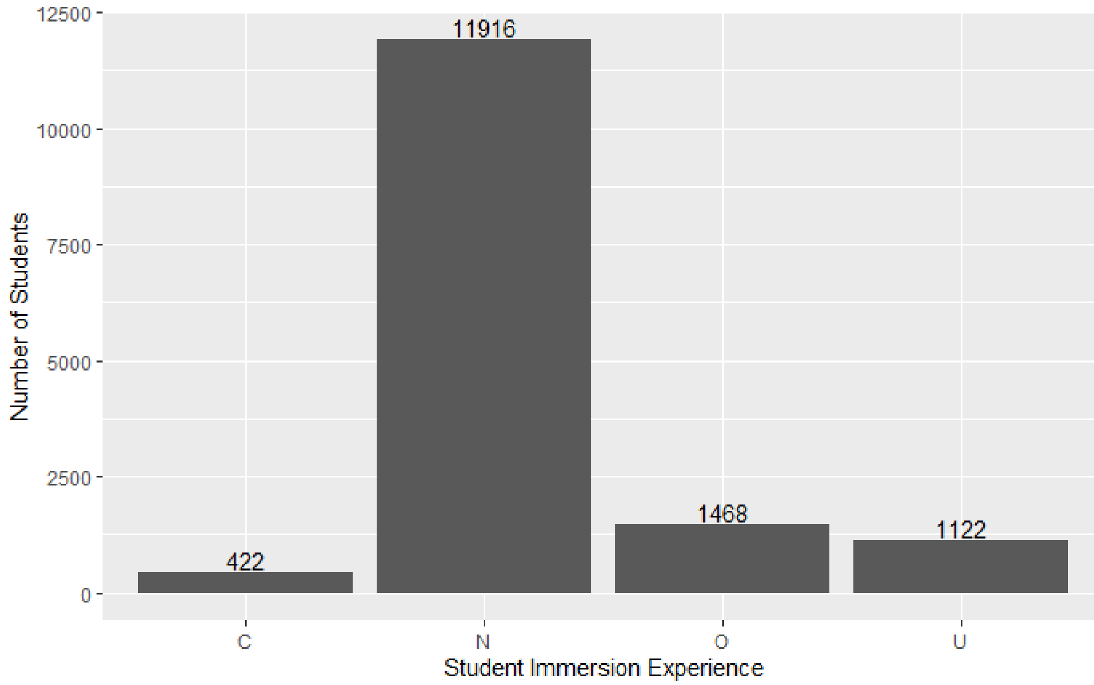


Table 14. Contingency Table for Student Immersion Experience

Immersion Experience	Total	Pass	Fail	Prop	PassRate	PropPass
N-No Immersion Experience	11916	4114	7802	79.82%	34.50%	70.78%
C-CONUS Immersion	422	247	175	2.83%	58.50%	4.25%
O-OCONUS Immersion	1468	812	656	9.83%	55.30%	13.97%
U-Unknown location Immersion	1122	639	483	7.52%	57.00%	10.99%
Overall	14928	5812	9116	100.00%	38.90%	100.00%

12. Student Distribution by Accumulated Years of Military Service

Figure 16 shows the student distribution by accumulated years of military service when the student started studying at DLIFLC. The vast majority of students have only one year of military experience or less. The data is strongly right-tailed with a second peak appearing at the six-year point.

Table 15 shows the contingency table for the same distribution as Figure 16. It is apparent that students with one year of accumulated service time (second column depicted in Figure 16) perform significantly worse than students with other lengths of service. To facilitate statistical analysis the students with similarly behaving year groups were grouped together. For more information on this grouping refer to the Variable Transformation section of this chapter, Table 4, or Table 16.

Figure 16. Student Distribution by Accumulated Years of Military Service

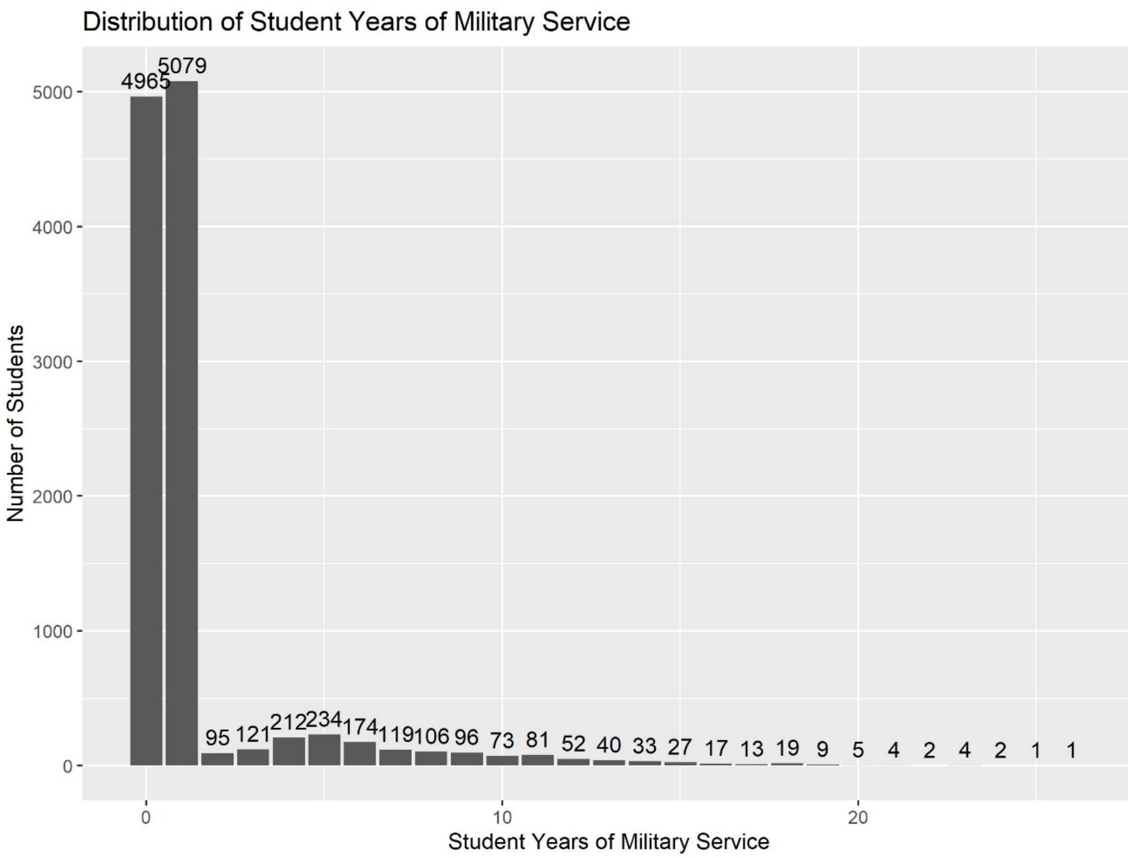


Table 15. Contingency Table for Years of Military Service (ungrouped)

Accumulated Years of Service at Arrival	Total	Pass	Fail	Prop	PassRate	PropPass
0	4965	2228	2737	42.900%	44.90%	47.37%
1	5079	1798	3281	43.800%	35.40%	38.23%
2	95	34	61	0.820%	35.80%	0.72%
3	121	45	75	1.040%	37.20%	0.96%
4	212	91	121	1.830%	42.90%	1.93%
5	234	116	116	2.020%	49.60%	2.47%
6	174	90	84	1.500%	51.70%	1.91%
7	119	46	72	1.030%	38.70%	0.98%
8	106	51	55	0.915%	48.10%	1.08%
9	96	43	53	0.829%	44.80%	0.91%
10	73	36	37	0.630%	49.30%	0.77%
11	81	39	42	0.699%	48.10%	0.83%
12	52	22	30	0.449%	42.30%	0.47%
13	40	16	24	0.345%	40.00%	0.34%
14	33	11	22	0.285%	33.30%	0.23%
15	27	10	17	0.233%	37.00%	0.21%
16	17	6	11	0.147%	35.30%	0.13%
17	13	5	8	0.112%	38.50%	0.11%
18	19	9	10	0.164%	47.40%	0.19%
19	9	1	8	0.078%	11.10%	0.02%
20	5	0	5	0.043%	0.00%	0.00%
21	4	3	1	0.035%	75.00%	0.06%
22	2	2	0	0.017%	100.00%	0.04%
23	4	0	4	0.035%	0.00%	0.00%
24	2	0	2	0.017%	0.00%	0.00%
25	1	1	0	0.009%	100.00%	0.02%
26	1	0	1	0.009%	0.00%	0.00%
Overall	11584	4703	6877	100.000%	40.60%	100.00%

Table 16. Contingency Table for Years of Military Service (grouped)

Accumulated Years of Service at Arrival	Total	Pass	Fail	Prop	PassRate	PropPass
0	4965	2228	2737	42.86%	44.90%	47.37%
1-3	5295	1877	3417	45.71%	35.40%	39.91%
4-6	620	297	321	5.35%	47.90%	6.32%
7	119	46	72	1.03%	38.70%	0.98%
8-11	356	169	187	3.07%	47.50%	3.59%
12-17	182	70	112	1.57%	38.50%	1.49%
18	19	9	10	0.16%	47.40%	0.19%
19-26	28	7	21	0.24%	25.00%	0.15%
Overall	11584	4703	6877	100.00%	40.60%	100.00%

13. Student Distribution by Motivation for Foreign Language Education

Figure 17 shows the student distribution in the SDB data set by self-reported personal motivation for study at DLIFLC. Students with motivation level 1 reported being "not motivated, does not want to study any language" on the initial student intake form. Motivation level 1 students made up 2% of the students. Less than 1% of students self-reported a motivation as a level of 2—"not motivated, prefers another language." 36% of students self-reported a motivation level of 3—"not my preferred language, but motivated to learn." 24% of students self-reported a motivation level of 4—"motivated, language is second or third choice." 37% of students self-reported a motivation level of 5—"motivated, language is first choice." The data is clearly left tailed, massed mostly on the right side of the graph with few students on the left. The majority of students report being content with the individual language of study. Table 17 shows the contingency table for student motivation level.

Figure 17. Student Distribution by Motivation

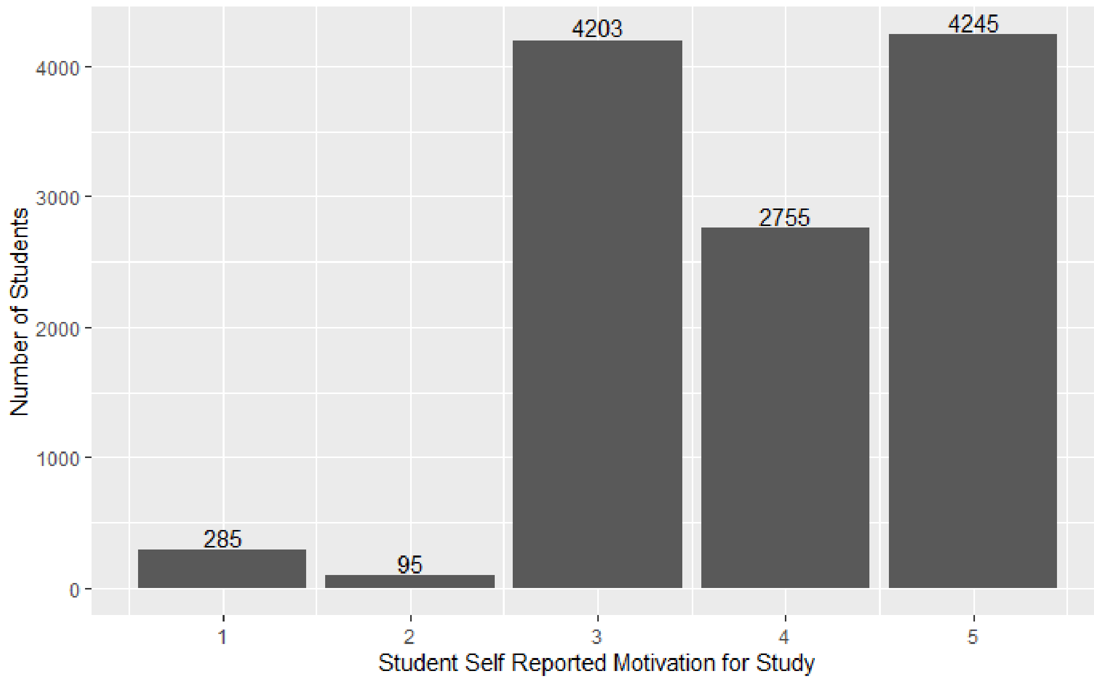


Table 17. Contingency Table for Student Motivation

Student Motivation	Total	Pass	Fail	Prop	PassRate	PropPass
1 (not motivated, doesn't wish to study a language)	285	115	170	2.46%	40.40%	2.45%
2 (not motivated, prefers a different language)	95	38	57	0.82%	40.00%	0.81%
3 (not preferred language, motivated to learn)	4203	1634	2569	36.29%	38.90%	34.74%
4 (motivated, language 2nd or 3rd choice)	2755	1085	1670	23.78%	39.40%	23.07%
5 (motivated, language 1st choice)	4245	1832	2413	36.65%	43.20%	38.95%
Overall	11583	4704	6879	100.00%	40.60%	100.00%

14. Student Distribution by DLAB Waiver

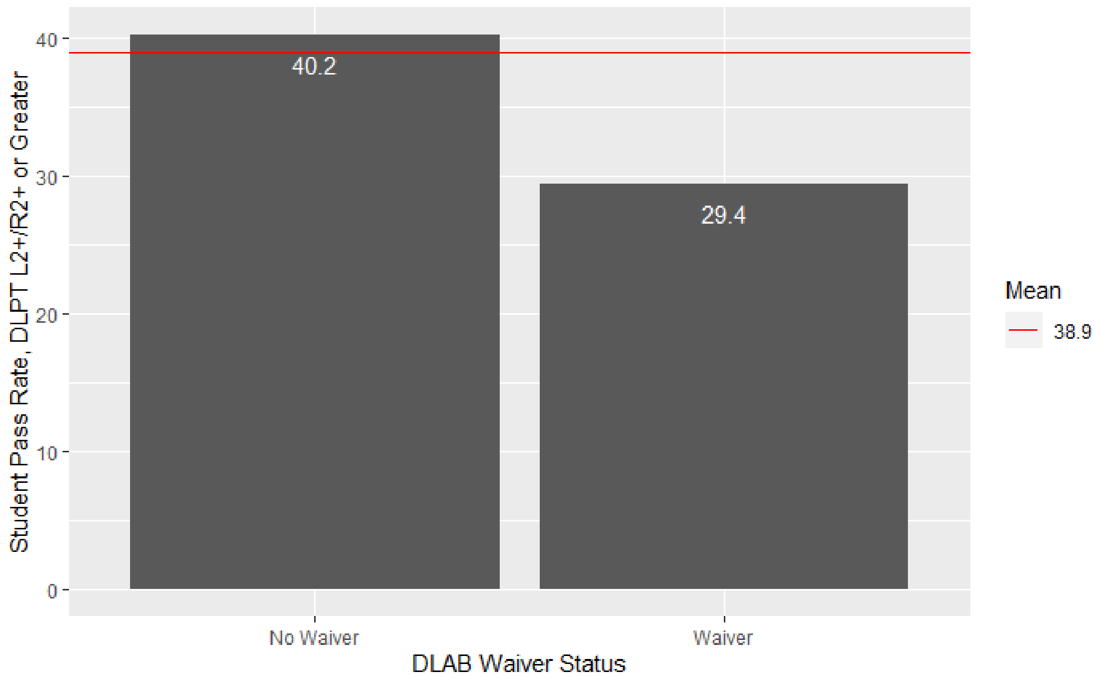
Table 18 shows the contingency table for students DLAB waiver status. Twelve percent of the student population received a DLAB waiver to enroll in a language that the student did not score high enough to enroll in.

The results of the chi-squared goodness of fit tests for the use of DLAB waivers has a *p*-value of approximately zero. A helpful visualization of the proportions of pass/fails is provided in Figure 18. The red line shows the overall pass rate of 38.9 –the average of the two group pass rates, weighted by the number of individuals in each group. We are able to reject the null hypothesis that there is no difference between the DLAB waiver statuses for students passing the DLPT at the L2+/R2+ level or greater. Students not utilizing DLAB waivers have historically higher rates of achieving L2+/R2+ or greater on the DLPT.

Table 18. Contingency Table for Student DLAB Waiver

DLAB Waiver Used?	Total	Pass	Fail	Prop	PassRate	PropPass
No	13144	5288	7856	88%	40.20%	90.98%
Yes	1784	524	1260	12%	29.40%	9.02%
Overall	14928	5812	9116	100%	38.90%	100.00%

Figure 18. Column Chart Depicting Student Performance Based on DLAB Waiver



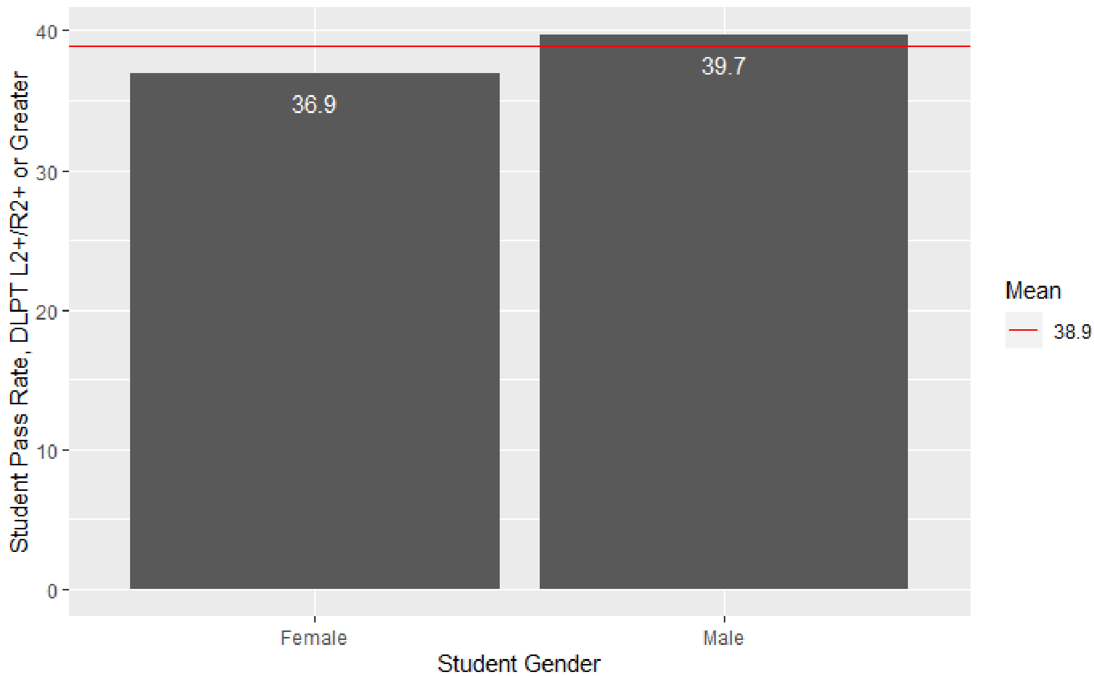
15. Student Distribution by Gender

Table 19 displays the student distribution table with respect to gender. Females make up 27% of the students, and males make up the other 73%. We are able to reject the null hypothesis that the two genders perform the same. Males performed significantly better (*p*-value < .01). This finding contradicts previous research conducted by Wong (2004). Figure 19 is a helpful depiction of the proportional passing rates of males and females.

Table 19. Contingency Table for Student Gender

Student Gender	Total	Pass	Fail	Prop	PassRate	PropPass
Female	4125	1524	2601	27.60%	36.90%	26.20%
Male	10803	4288	6515	72.40%	39.70%	73.80%
Overall	14928	5812	9116	100.00%	38.90%	100.00%

Figure 19. Column Chart Depicting Student Performance Based on DLAB Waiver



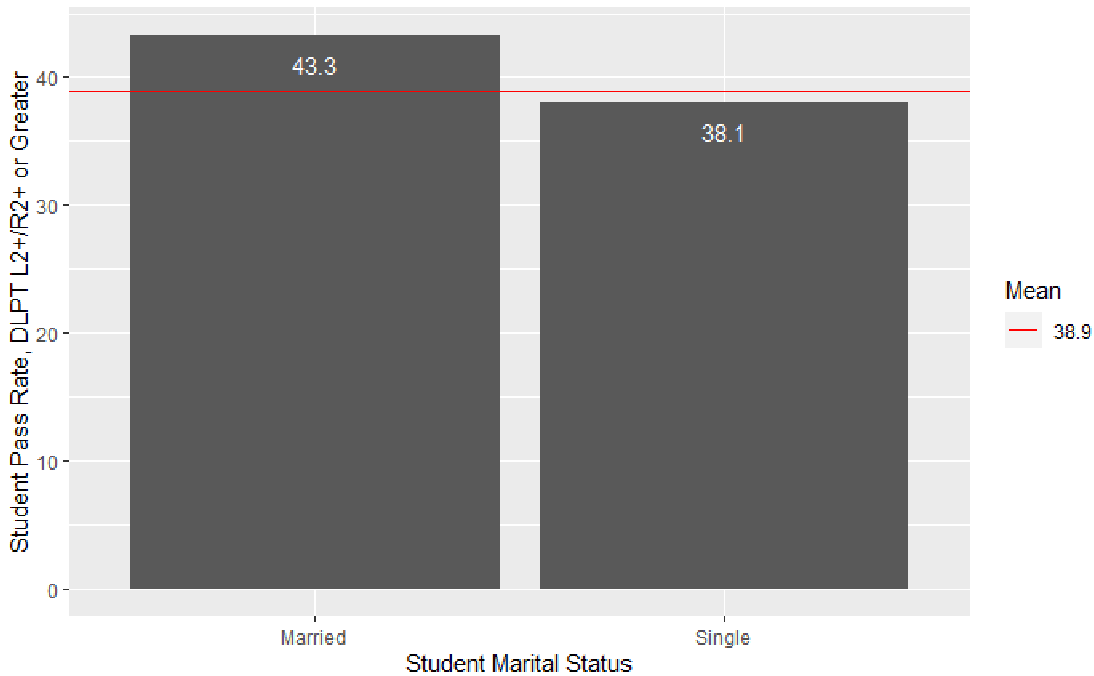
16. Student Distribution by Marital Status

Table 20 shows the distribution of students with respect to individual marital status. Married students make up 16% of the total student population. The results of the chi-squared goodness of fit tests for student's marital status indicated a *p*-value of nearly zero. A helpful visualization of the proportions of pass/fails is provided in Figure 20. We were able to reject the null hypothesis that there is no difference between the marital statuses for students passing the DLPT at the L2+/R2+ level or greater. Married students have historically higher rates of students achieving L2+/R2+ or greater on the DLPT.

Table 20. Contingency Table for Student Marital Status

Student Marital Status	Total	Pass	Fail	Prop	PassRate	PropPass
Married	2414	1046	1368	16.20%	43.30%	18%
Single	12513	4765	7748	83.80%	38.10%	82%
Overall	14927	5811	9116	100.00%	38.90%	100%

Figure 20. Column Chart Depicting Student Performance Based on Marital Status



THIS PAGE INTENTIONALLY LEFT BLANK

IV. ANALYSIS AND RESULTS

This chapter reviews the diagnostics and results of the four models created using the SDB data set to predict a student's likelihood of success of passing the DLPT at the 2L+/2R+ level or greater. There were two distinct types of models built: random forests and neural networks. Each model was first built using the predictors in Bermudez-Mendez's logistic regression model. The approach allowed our team to draw direct comparisons between the models. We explored each model type for its respective predictor optimization once the direct comparisons were complete.

We examined the student's respective language backgrounds to answer questions like: "Does a student with a Chinese background have a better chance of passing the DLPT at the 2L+/2R+ level or higher when learning Korean than one with a Spanish background?" We also examined other predictors to see what other associations exist for a student's chances of passing the DLPT. The other predictors that were statistically assessed were: DLAB waiver, and years of military service, service branch, gender, motivation for study, marital status, prior language source, prior language experience, prior language proficiency, induction status, and level of education. These analyses were performed with chi-squared tests in the hopes of finding potential policy recommendations to improve DLIFLC's ability to graduate more successful students.

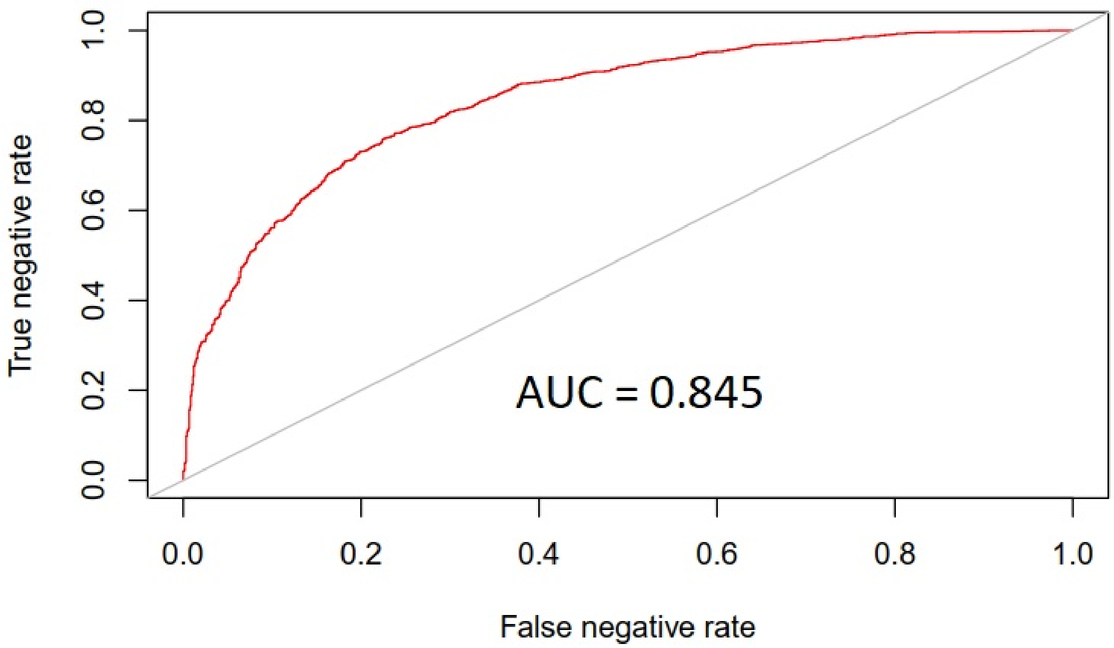
A. RANDOM FOREST MODEL WITH BERMUDEZ-MENDEZ PREDICTORS

1. Goodness of Fit

We first assessed each model's goodness of fit. A model with an AUC from 0.8 to 0.9 is considered exceptional at discerning between failure and success (Hosmer 2013, p. 177). The random forest model built with the predictors Bermudez-Mendez used in his logistic regression 3rd semester model study has an AUC of 0.845 when applied to the test set. Our model can be considered excellent at discriminating between successes and failures.

The random forest model outperforms the Bermudez-Mendez logistic regression 3rd semester model which achieved an AUC of 0.838 when applied to the test set. Figure 21 shows the graphical representation of the AUC curve for the random forest model. For direct comparison to the Bermudez-Mendez 3rd semester model outputs, refer to Appendix F.

Figure 21. AUC Curve for Random Forest Model Built with Bermudez-Mendez Predictors



2. Classification Table

After the random forest model was made using Bermudez-Mendez's predictors, we used a test set made up of the last 25% of the data to evaluate its performance. The results of this were tabulated into a classification table that can be seen in Table 21. The derived statistics from the classification table can be seen in Table 22.

The overall accuracy of 0.77 in Table 22 means that we can correctly predict a success or failure at a rate of 77%. This can be considered a good result, based on how a naive person would guess "failure" every time with approximately 61% accuracy. The sensitivity of 0.84 means that we can predict success for a successful student correctly 84% of the time. The specificity of 0.67 means that we correctly predict failure for a student that fails to achieve L2+/R2+ or greater on the DLPT 67% of the time. The PPV tells us that when we predict a student to be successful we are correct 79% of the time. Similarly, the NPV tells us that when we predict a student to fail, there is a 77% chance the student actually will.

Table 21. Classification Table for Random Forest Model with Bermudez-Mendez Predictors

Predicted/Observed	Success	Failure	
Success	**1438**	**277**	1715
Failure	**389**	**783**	1172
	1827	1060	

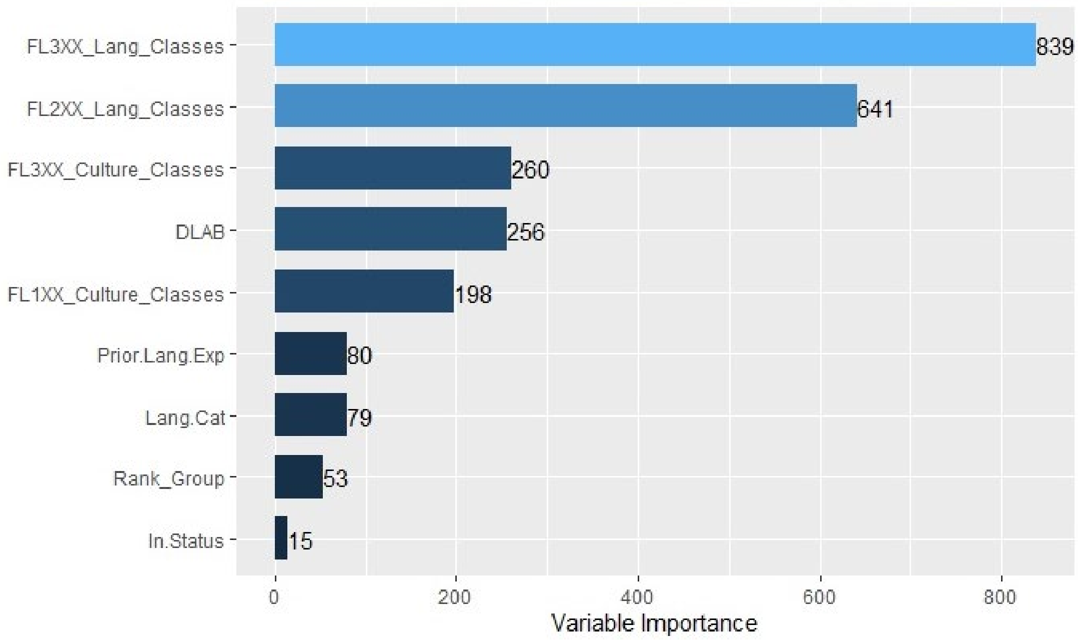
Table 22. Random Forest Model with Bermudez-Mendez Predictors Classification Derived Metrics

Sensitivity	0.84
Specificity	0.67
Precision (PPV)	0.79
(NPV)	0.74
Overall Accuracy	0.77
AUC:	0.845

3. Variable Importance

In Figure 22 the importance of the Bermudez-Mendez selected predictors used to make the random forest model are shown, and ranked by respective importance. The random forest model assigned different weights (relative importance) to the predictors than the logistic regression used by Bermudez-Mendez. The table for the variable importance associated with the Bermudez-Mendez 3rd semester logistic regression model can be found in Appendix E.

Figure 22. Variable Importance for Random Forest Model with Bermudez-Mendez Predictors

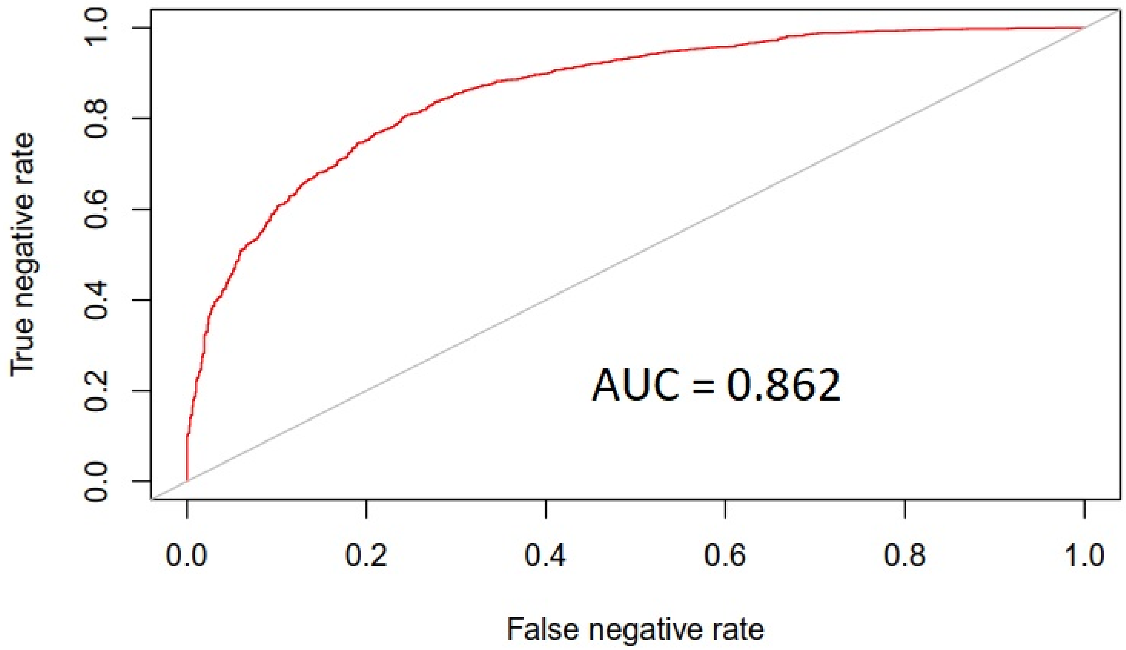


B. RANDOM FOREST MODEL WITH PREFERRED PREDICTORS

1. Goodness of Fit

The random forest model that was allowed to select the best mix of predictors performed the best out of the four models that were generated in this study. This random forest model achieved an AUC of 0.862 when applied to the test set, nearly 2% better than the previous random forest model. The AUC curve for the RF model can be seen in Figure 23.

Figure 23. AUC Curve for Random Forest Model Built with Preferred Predictors



2. Classification Table

The random forest model that used optimized predictors achieved 2% better performance in overall accuracy than the previous random forest model at 0.79, also approximately 2% better than the Bermudez-Mendez logistic regression. This random forest model also achieved the best sensitivity of all the models at 0.85. These results are tabulated in Table 23 and Table 24.

Table 23. Classification Table for Random Forest Model with Preferred Predictors

Predicted/Observed	Success	Failure	
Success	**1458**	**257**	1715
Failure	**347**	**825**	1172
	1805	1082	

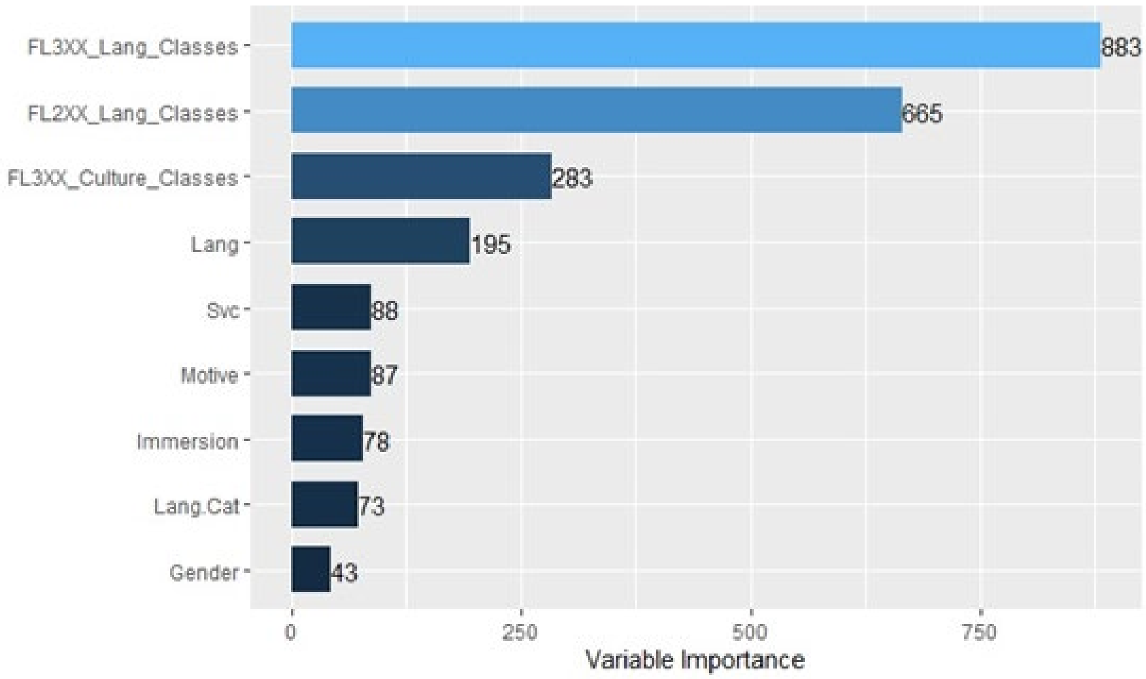
Table 24. Random Forest Model with Preferred Predictors Classification Derived Metrics

Sensitivity	0.85
Specificity	0.70
Precision (PPV)	0.81
(NPV)	0.76
Overall Accuracy	0.79
AUC:	0.862

3. Variable Importance

To build the random forest model with optimized predictors, it was first built in the Ranger RF package (Wright and Zeigler 2017) using all available predictors and hyper-tuned using a Latin Hypercube Sampling (LHS) of different parameter tunings. The random forest was then subjected to a backward elimination predictor selection routine to minimize overfitting of the model. Figure 24 shows the variables selected for the best random forest model. The variable optimized random forest kept the 3rd semester model predictors for the language classes, culture classes, and the language category. The random forest model selected some new predictors: language studied, service branch, motivation, immersion status, and gender.

Figure 24. Variable Importance for Random Forest Model with Preferred Predictors



C. NEURAL NETWORK MODEL WITH BERMUDEZ-MENDEZ PREDICTORS

1. Goodness of Fit

A neural network model was built using the Bermudez-Mendez 3rd semester model predictors for accurate comparison to the other models. In all, it achieved an AUC of 0.837. Like all the other models in this study it was hyper-tuned using a LHS method to adjust the tuning parameters in H2o NN package (LeDell 2020). This result is marginally better than the original 3rd semester model, which had an AUC of 0.833 when applied to the test set.

2. Classification Table

The classification table for the NN utilizing Bermudez-Mendez 3rd semester model predictors can be seen in Table 25 and the derived statistics are visualized in Table 26. Overall the results of this NN are underwhelming. It actually had worse performance in overall accuracy than the 3rd semester logistic regression model. The one bright spot is this model produced the best specificity of all 5 models a score of 0.76. The specificity of 0.76

means that we correctly predicted failure for a student who could not achieve L2+/R2+ or greater on the DLPT 76% of the time.

Table 25. Classification Table for Neural Network Model with Bermudez-Mendez Predictors

Predicted/Observed	Success	Failure	
Success	**2930**	**901**	3831
Failure	**1469**	**4644**	6113
	4399	5545	

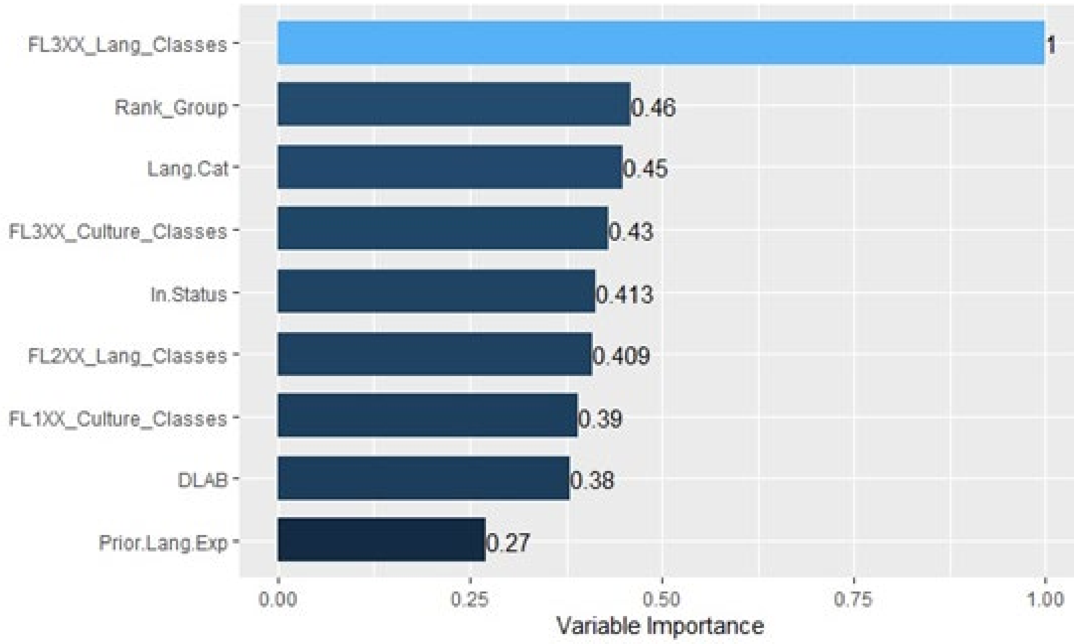
Table 26. Neural Network Model with Bermudez-Mendez Predictors Classification Derived Metrics

Sensitivity	0.76
Specificity	0.76
Precision (PPV)	0.67
(NPV)	0.84
Overall Accuracy	0.76
AUC:	0.837

3. Variable Importance

In Figure 25 are the tabulated weights for the predictors in the NN model using the Bermudez-Mendez 3rd semester logistic regression predictors. It is interesting to note that like the RF model, the NN assigned very different weights to the predictors. A direct comparison in the weighting of the factors in the 3rd semester logistic regression model can be found in Appendix E.

Figure 25. Variable Importance for Neural Network Model with Bermudez-Mendez Predictors



D. NEURAL NETWORK MODEL WITH PREFERRED PREDICTORS

1. Goodness of Fit

The NN model utilizing optimized predictors produced an AUC of 0.845 when applied to the test set, approximately 1% better than the Bermudez-Mendez 3rd semester logistic regression model. It should be noted that the science of hyper-tuning NN models is still not as developed as with other modeling forms, and there remains some more potential for improvement in the model. This model was hyper-tuned using a Latin Hypercube Sampling (LHS) of different parameter tunings. The NN was then subjected to a backward elimination predictor selection routine to minimize model overfitting.

The difficulty with NNs is that their model output is somewhat stochastic in nature and the AUC or accuracy can sway up and down across a couple percentage points across iterations. This can make it problematic in nailing down an improving hyperparameter tuning. Improvements in tuning can be potentially missed if the random variability is not working in the model’s favor during the random sampling round.

2. Classification Table

The classification table for the predictor optimized NN is displayed in Table 27, and the derived metrics are displayed in Table 28. This model did not stand out in any particular way other than 1) having a better AUC than the Bermudez-Mendez 3rd semester logistic regression model, and 2) having the best NPV score of all the models with a score of 0.84.

Table 27. Classification Table for Neural Network Model with Preferred Predictors

Predicted/ Observed	**Success**	**Failure**	
Success	**2844**	**713**	3557
Failure	**1338**	**3820**	5158
	4182	4533	

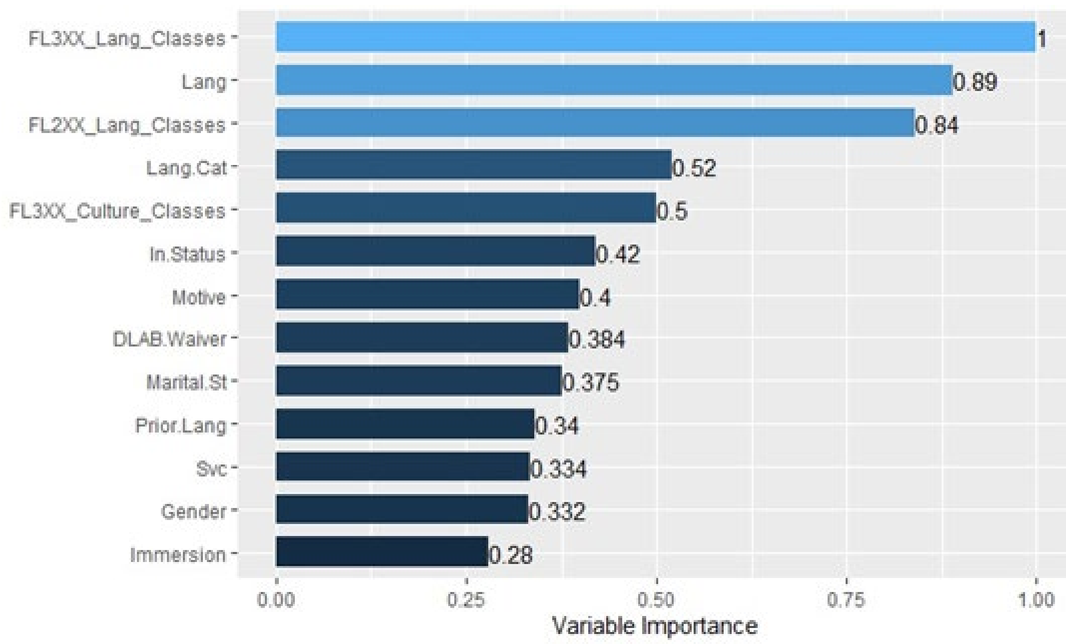
Table 28. Neural Network Model with Preferred Predictors Classification Derived Metrics

Sensitivity	0.80
Specificity	0.74
Precision (PPV)	0.68
(NPV)	0.84
Overall Accuracy	0.76
AUC:	0.845

3. Variable Importance

To build the NN model with optimized predictors, it was first built in the H2o package (Cook 2017) using all available predictors and hyper-tuned using a Latin Hypercube Sampling (LHS) of different parameter tunings. The NN was then subjected to a backward elimination predictor selection routine to minimize overfitting. Figure 26 shows the predictors that were chosen by the NN to maximize its accuracy. Overall the only predictors it keeps in common with the Bermudez-Mendez 3rd semester model predictors are the language classes, culture classes, the induction status, and the language category. The NN model was optimized by also selecting many new predictors to include: language studied, service branch, motivation, prior foreign language, marital status, immersion status, and gender.

Figure 26. Variable Importance for Neural Network Model with Preferred Predictors



E. OVERALL MODEL COMPARISON

The overall model comparison is tabulated in Table 29. The best numbers produced by the various models are in bold typeface. The best model in terms of overall performance was the RF model using optimized predictors. It should be noted that both the RF and NN models in all cases produced nearly equal or superior results to the logistic regression.

Table 29. Overall Model Comparison

	Bermudez-Mendez Logistic Regression 3rd Semester Model	Bermudez-Mendez Predictors Random Forest	Bermudez-Mendez Predictors Neural Network	Preferred Predictors Random Forest	Preferred Predictors Neural Network
Sensitivity	0.802	0.838	0.765	**0.850**	0.800
Specificity	0.706	0.668	**0.760**	0.704	0.741
Precision (PPV)	**0.816**	0.787	0.666	0.808	0.680
(NPV)	0.686	0.739	0.838	0.762	**0.843**
Overall Accuracy	0.765	0.769	0.762	**0.791**	0.765
AUC:	0.838	0.845	0.837	**0.862**	0.845

F. STATISTICAL ANALYSIS OF LANGUAGE BACKGROUNDS

1. Statistical Results of Prior Language

Multivariate models are the best method to analyze complex interactions in data sets and predict outcomes, but these models have the significant limitation that their inner workings are difficult to understand. Univariate analysis has the advantage in being relatively easy to understand, and can help find abnormalities in the data. The rest of the results presented in Chapter 5 will be univariate analysis using chi-squared test statistics shown in tabulated form. The chi-squared test was performed for each language background against the null value of no language background. At the end of each table the null value of no prior language background was compared against all other language backgrounds within the studied group.

The following is the methodology of how the results are presented for the rest of this chapter: Variables that did not meet the criterion of having at least five students expected to pass under the test's hypothesis have the number of students highlighted in red. The results in each table will be sorted by their proportion of student success from high to low. The proportion is simply the number of students who achieved L2+/R2+ or greater on the DLPT divided by all students in that particular category. The larger the proportion, the better that group of students did. There is a line drawn below indicating where the cutoff for the mean proportion of students for that language occurred. A column chart will also be presented for each predictor graphically displaying the proportion of success that occurred. This chart will also contain a line depicting where the mean performance was for that specific view of the data set.

Many of the languages tested displayed striking results in terms of proportions of students that passed for a specific language taught at DLIFLC. Unfortunately many of these languages did not have enough students to meet the minimum criterion of at least five students predicted to pass to meet the assumptions for using a chi-squared table for statistical inference. In many cases, even if enough students were present to use the test, not enough existed to drive enough statistical power for a significant level of *p*-value. These languages will hopefully receive further study in the future when there is more student data available. For instance Greek-Oghuz language group students performed extremely well (0.71 student pass proportion, versus a mean of 0.26) on the Spanish DLPT, but did not have enough students to meet the minimum number of students to use the chi-square test. Only language backgrounds that met the minimum assumptions for using the chi-square test and have significant *p*-value will appear in Table 30, the summary table of the language background study. In order to see all the background language results refer to Appendix F.

Table 31 and Figure 27 show that students with no foreign language background (English only speakers), students grouped into the "other" language group, and Spanish speakers did not fare well when viewed across all the different languages taught at DLIFLC. Spanish is special in that it is the only language where students with prior experience in it fared worse than the average student taking Spanish. English-only speakers were only able to score above the mean when taking Persian-Farsi or Hebrew. Refer to Appendix F for the data tables displaying these results. Table 31 and Figure 27 also show the language groups for Korean, Hebrew, Greek-Oghuz, Chinese, Arabic, Austronesian, French, German, and Romance performed extremely well across all languages.

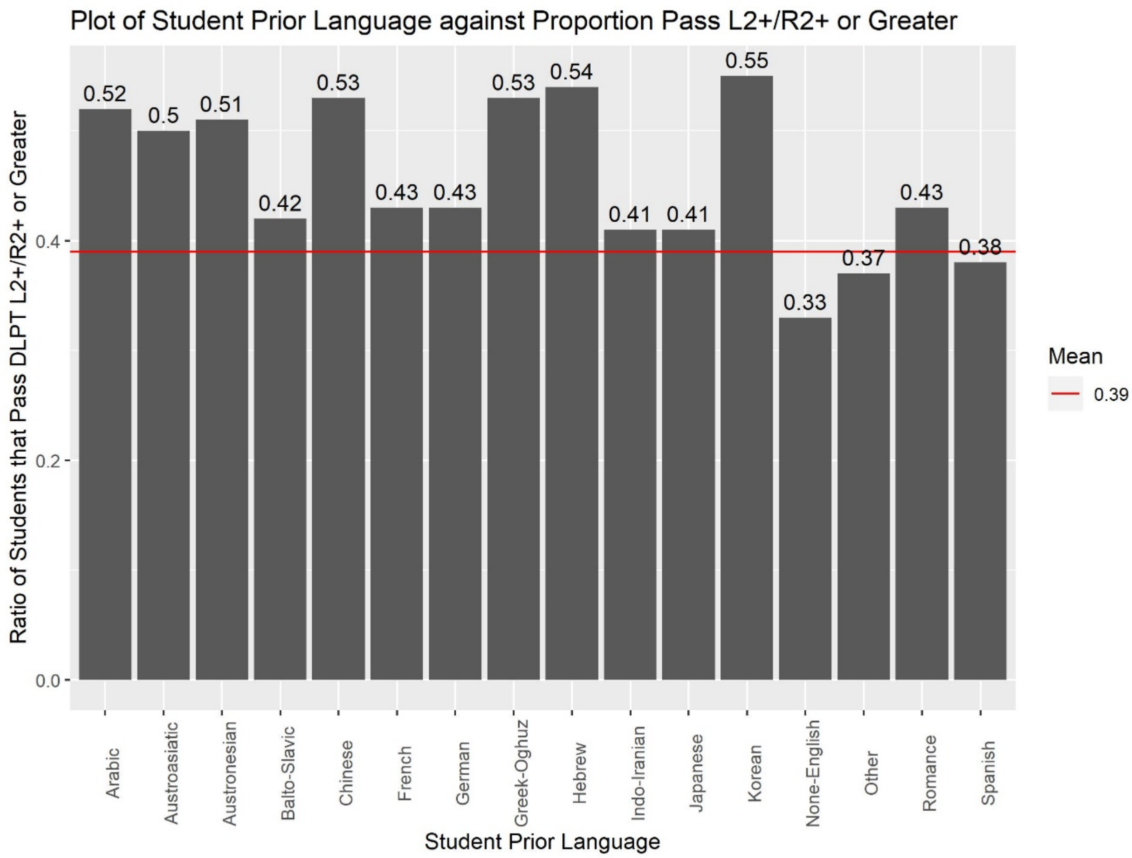
Table 30. Summary Table of Results with Significant *p*-value for Language Background Study

Languages Taught at DLIFLC	Number of Students	Background Language Performance Above The Mean	Background Language Performance Below the Mean
Arabic (AD, AE, AP, DG) Combined	3995	Korean, French, German	None-English
Chinese-Mandarin	2010	Chinese	NA
French	509	Arabic, French, German	None-English
Indonesian	289	Spanish	None-English
Korean	1437	Korean, Romance	None-English
Persian-Farsi	2142	Indo-Iranian	Romance
Pashto (Pashtu-Afghan)	1065	Greek-Oghuz, Indo-Iranian, Arabic, French, Chinese, Romance, Spanish	German, Other, None-English
Spanish	1488	Greek-Oghuz, Austronesian, Indo-Iranian, Arabic	NA
Russian	1294	Japanese, Chinese, Balto-Slavic	None-English

Table 31. Overall Language Background Groups Performance Best to Worst

Prior Language Group	Total Students	Pass	Fail	Proportion Pass	Chi-sq *p*-value
Korean	206	114	92	0.55	~0
Hebrew	39	21	18	0.54	< .05
Greek-Oghuz	58	31	27	0.53	< .01
Chinese	306	162	144	0.53	~0
Arabic	317	166	151	0.52	~0
Austronesian	115	59	56	0.51	~0
Austroasiatic	40	20	20	0.50	< .05
French	1121	483	638	0.43	~0
German	852	365	487	0.43	~0
Romance	529	226	303	0.43	~0
Balto-Slavic	295	125	170	0.42	< .01
Indo-Iranian	131	54	77	0.41	> .05
Japanese	159	65	94	0.41	> .05
Spanish	4313	1660	2653	0.38	~0
Other	3102	1150	1952	0.37	< .01
None-English against all others	3349	1110	2239	0.33	~0
Total:	14932	5811	9121	0.39	

Figure 27. Plot of Student's Prior Language Group Performance across All Languages



2. Results for Student's Prior Language Source

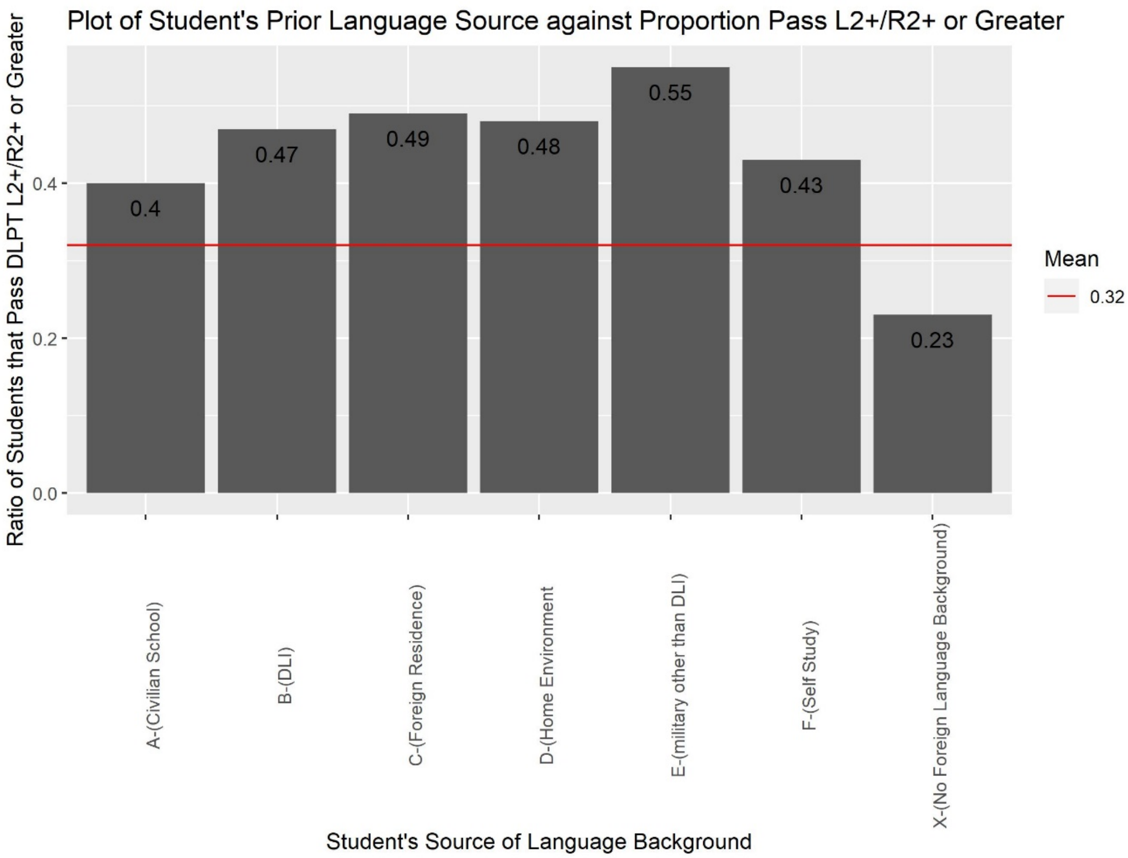
The methodology of studying the proportional pass rates for the predictor's prior language source, experience, and proficiency are done a little differently than the other investigated predictors. In each case the prior language source/experience/proficiency was first tested for all students across all language backgrounds. Next these same tests were conducted for students with French and Spanish backgrounds to see if there were differences between categories within these languages. Spanish was chosen for study because it has the largest number of students and they showed generally inferior behavior when compared to other language backgrounds (see Table 31 and Figure 27). French was the other background chosen for study because: 1) it was the second largest language group background, and 2) unlike Spanish, its students showed relatively good performance across the different languages taught at DLIFLC (see Table 31 and Figure 27).

The results of the chi-squared goodness of fit tests for the student's prior language source across all language backgrounds are shown in Table 32. A helpful visualization of the proportions of pass/fails is provided in Figure 28. Each of the different possible language sources were compared to the null value of X (NA) individual in two-way tables. For completeness the X (NA) value was compared to all the other language sources grouped together to form a single two-way comparison. We are able to reject the null hypothesis that there is no difference between the sources of language backgrounds for students passing the DLPT at the L2+/R2+ level or greater. Each language source was statistically significantly different from the X group. Students with a language background (other than none) performed significantly higher than the mean.

Table 32. Student's Prior Language Source Overall Chi-Squared Results

Student Prior Language Source (all students)	**Total Students**	**Pass**	**Fail**	**Proportion Pass**	**Chi-sq *p*-value**
E (Military – Other than DLI)	91	50	41	0.55	~ 0
C (Foreign Residence)	626	304	322	0.49	~ 0
D (Home Environment)	666	317	349	0.48	~ 0
B (DLI)	712	334	378	0.47	~ 0
F (Self Study)	269	116	153	0.43	~ 0
A (Civilian School)	6267	2492	3775	0.40	~ 0
X (NA) - against all others	9742	2199	7543	0.23	~ 0
Total:	18373	5812	12561	0.32	

Figure 28. Column Chart Depicting Student Performance Overall Based on Prior Language Source



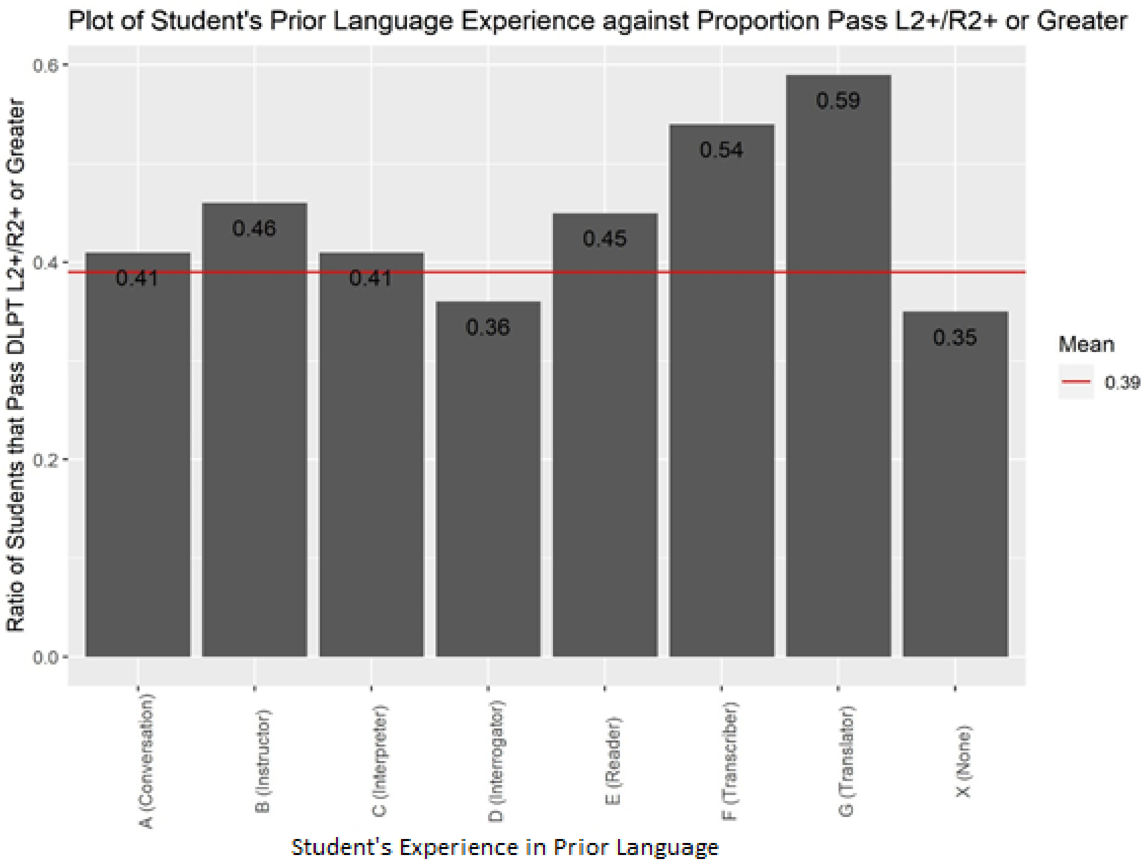
Perhaps surprisingly, there was no significant difference in the rate at which students with prior French experience passed the DLPT at the L2+/R2+ or greater level associated with differing levels of prior language source. The power of this test may be small due to the fact that about 89% of these students were in the A (Civilian School) group. The story is different with students with prior Spanish: the groups are statistically different overall ($\chi^2 = 20.5$ on 5 d.f., $p \approx 0$), driven by the fact that those in the A group were less likely to reach the L2+/R2+ level than others ($\chi^2 = 6.1$ on 1 d.f., $p = 0.013$).

The results of the chi-squared goodness of fit tests for the student's prior language experience across all language backgrounds are shown in Table 33. A helpful visualization of the proportions of pass/fails is provided in Figure 29. We are able to reject the null hypothesis that there is no difference between the experiences of language backgrounds for student's historical performance in passing the DLPT at the L2+/R2+ level or greater. The set of experience levels is adjudged different, and, individually, each language experience except for interrogator and interpreter, whose samples were small, were significantly different from the "no language" group. Students with no language background performed significantly below the mean success rate.

Table 33. Student's Prior Language Experience Overall Chi-Squared Results

Student Prior Language Experience (all students)	**Total Students**	**Pass**	**Fail**	**Proportion Pass**	**Chi-sq *p*-value**
G (Translator)	228	134	94	0.59	~ 0
F (Transcriber)	197	106	91	0.54	~ 0
B (Instructor)	585	270	315	0.46	~ 0
E (Reader)	656	295	361	0.45	~ 0
A (Conversation)	5543	2300	3243	0.41	~ 0
C (Interpreter)	101	41	60	0.41	> .05
D (Interrogator)	22	8	14	0.36	> .05
X (None) - against all others	7599	2656	4943	0.35	> .05
Total:	14931	5810	9121	0.39	

Figure 29. Column Chart Depicting Student Performance Overall Based on Prior Language Experience



As with the source of language, the various levels of experience were not a good predictor of DLPT success for students with French backgrounds. With Spanish, as with French, some categories were sparse, but comparing the most populated ones showed no real difference across the various levels of experience.

3. Results for Student's Prior Language Proficiency

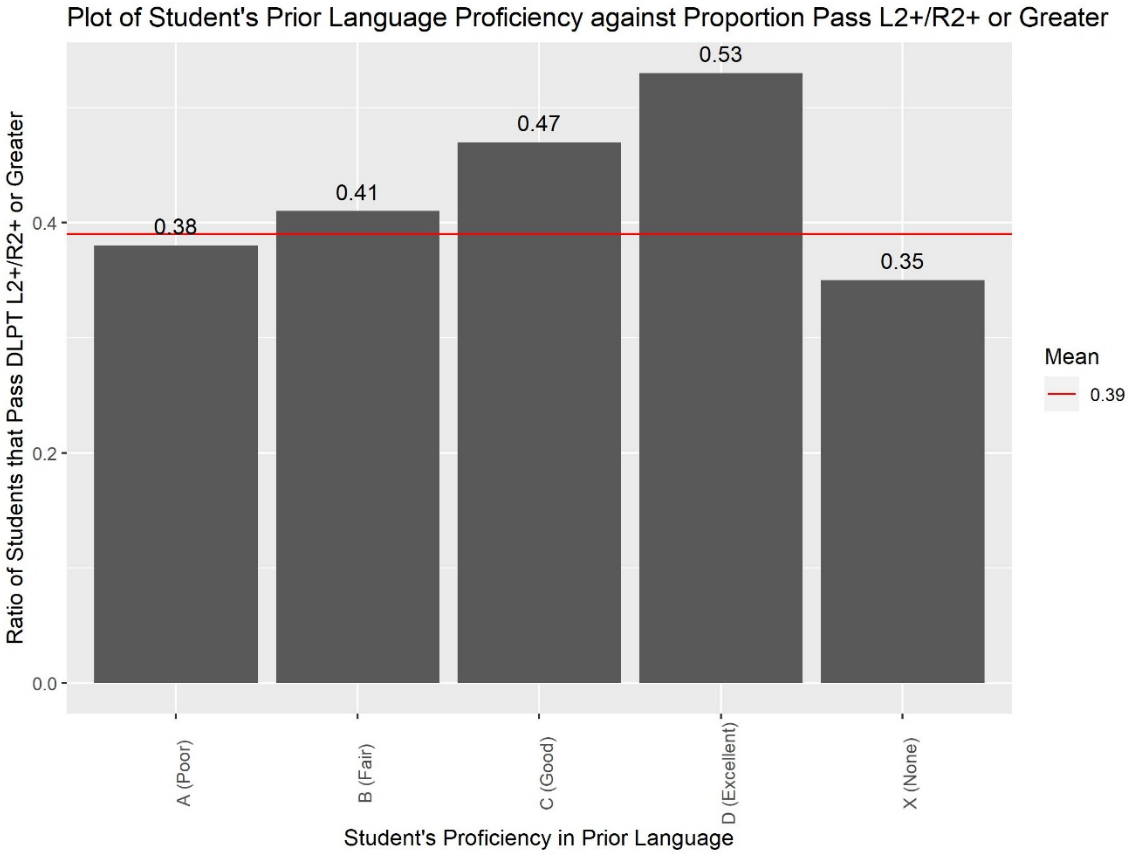
The results of the chi-squared goodness of fit tests for the student's self-reported prior language proficiency across all language backgrounds are shown in Table 34. A helpful visualization of the proportions of pass/fails is provided in Figure 30. We are able to reject the null hypothesis that there is no difference between the proficiency levels of language backgrounds for student's historical performance in passing the DLPT at the

L2+/R2+ level or greater ($\chi^2 = 155$, 4 d.f., $p \approx 0$). All language skill levels showed statistically significant passing rates individually against the “None” group.

Table 34. Student’s Prior Language Proficiency Overall Chi-Squared Results

Student Prior Language Source (all students)	**Total Students**	**Pass**	**Fail**	**Proportion Pass**	**Chi-sq *p*-value**
D (Excellent)	669	355	314	0.53	~ 0
C (Good)	1569	738	831	0.47	~ 0
B (Fair)	2855	1182	1673	0.41	~ 0
A (Poor)	3192	1222	1970	0.38	< .01
X (None) - against all others	6648	2315	4333	0.35	~ 0
Total:	14933	5812	9121	0.39	

Figure 30. Column Chart Depicting Student Performance Overall Based on Prior Language Proficiency



There is no discernible effect of proficiency among students with French background. The levels of proficiency were statistically separable among those with Spanish (χ = 31.7, 4 d.f., $p \approx 0$). Students with proficiency levels A (Poor) and X (None) had success rates substantially lower than members of other groups.

G. STATISTICAL ANALYSIS OF OTHER PREDICTORS

The same type of chi-squared statistical analysis that was performed on the prior language groups was also performed on the other predictors that were available in the SDB dataset. This work was intended to help recruiting commands shape their policies for what students the recruiters bring in and accept into their language training programs. By turning the results of these computations into policy there is a possibility that DLPT scores can be raised without any additional monetary investment by the U.S. military. In these cases all the levels are compared simultaneously. We do not also compare each levels against a baseline value, as with prior languages, because these predictors do not have a "none" category. Overall pass rates can change from table to table depending on missing values.

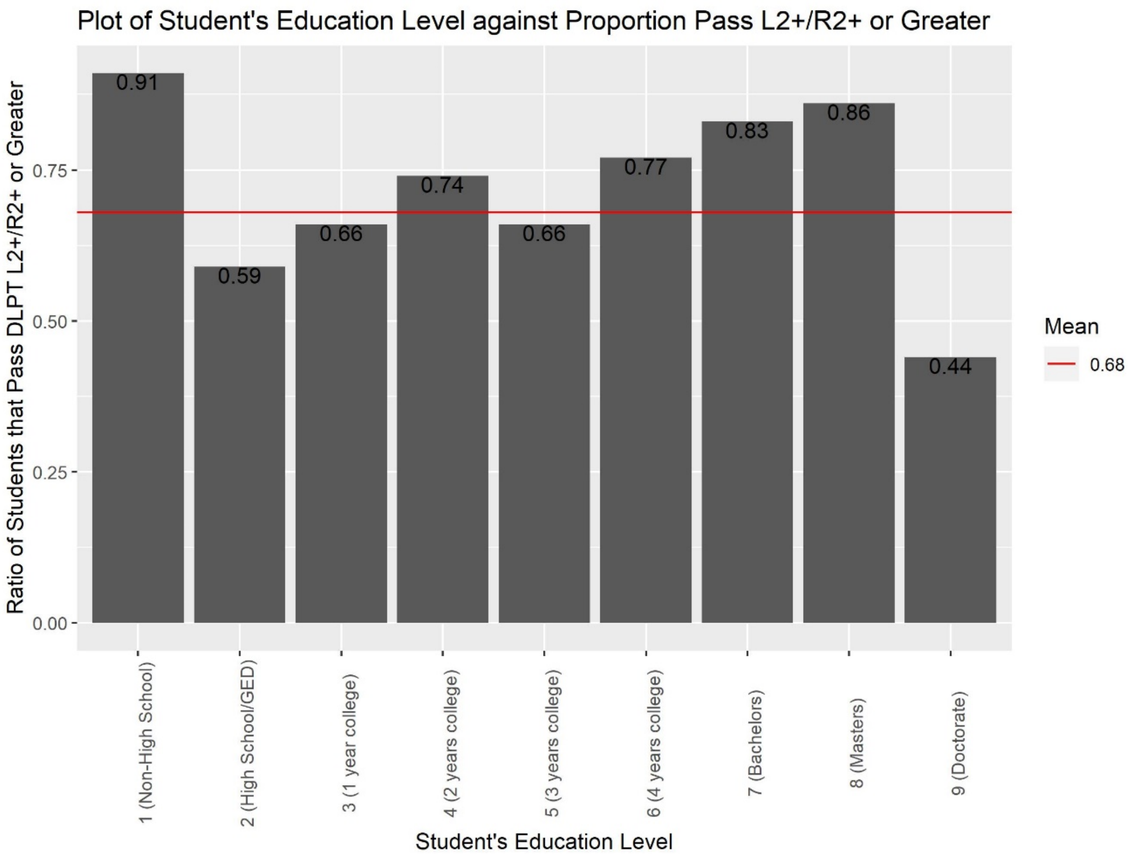
1. Results for Student's Education Level

The results of the chi-squared goodness of fit tests for the student's self-reported education level is shown in Table 35. A helpful visualization of the proportions of pass/ fails is provided in Figure 31. We are able to reject the null hypothesis that there is no difference between the education levels of statuses for passing the DLPT at the L2+/R2+ level or greater. Students with masters, bachelors, and (in small numbers) non-high school diplomas had the highest passing rates.

Table 35. Student's Education Level Results

Student Education Level	**Total Students**	**Pass**	**Fail**	**Proportion Pass**	**Chi-sq**
1 (Non-High School)	21	10	11	0.48	
8 (Masters)	350	162	188	0.46	
7 (Bachelors)	2321	1053	1268	0.45	
6 (4 years college)	424	184	240	0.43	
4 (2 years college)	1892	802	1090	0.42	
5 (3 years college)	751	299	452	0.40	
3 (1 year college)	1640	650	990	0.40	$\chi^2 = 57.6$
2 (High School/GED)	4112	1521	2591	0.37	9 d.f.
9 (Doctorate)	23	7	16	0.30	$p \approx 0$
Total:	11534	4688	6846	0.41	

Figure 31. Column Chart Depicting Student Performance Based on Education Level



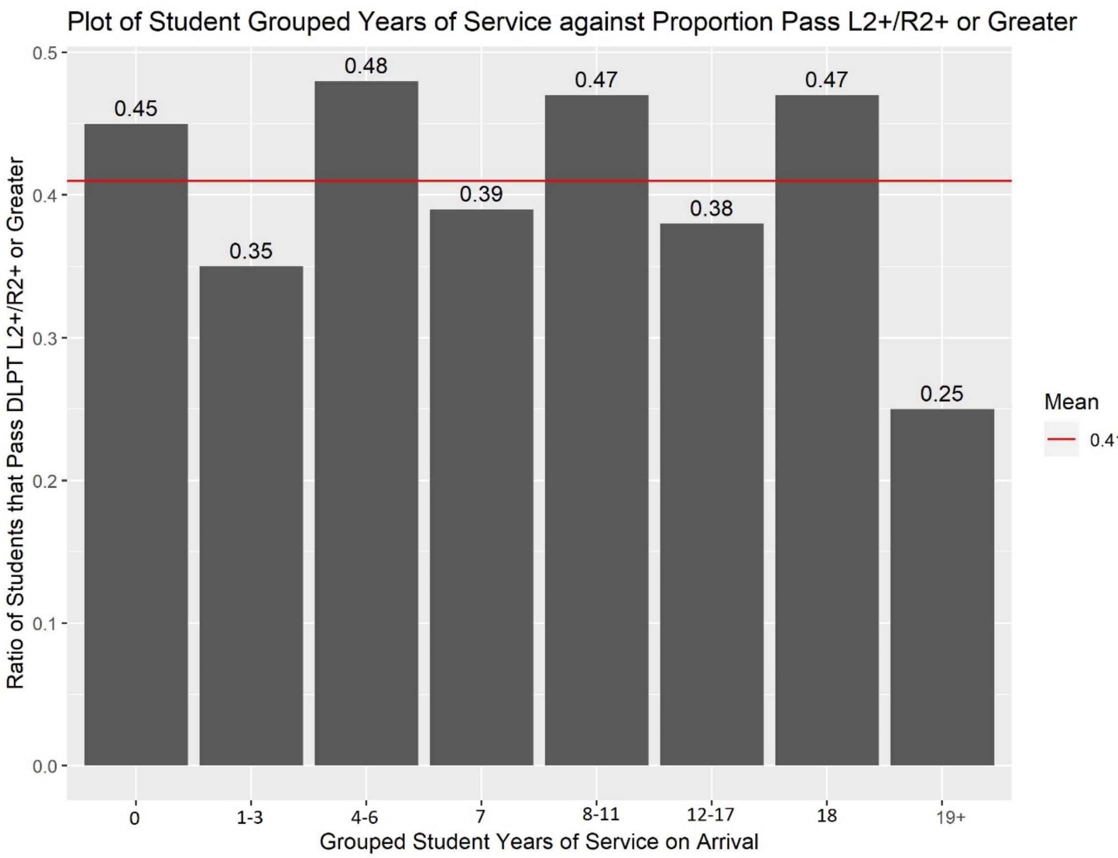
2. Results for Student's Years of Service

The results of the chi-squared goodness of fit tests for the student's accumulated years of service are shown in Table 36. A helpful visualization of the proportions of pass/ fails is provided in Figure 32. We are able to reject the null hypothesis that there is no difference among ages students passing the DLPT at the L2+/R2+ level or greater. Students with 0, 4–6 and 8–11 years of service of service had the highest passing rates. The dips in performance that occurs across the student year spectrum make it difficult to suggest policy fixes to select students likely to be successful.

Table 36. Student's Years of Service Chi-Squared Results

Accumulated Years of Service at Arrival	**Total Students**	**Pass**	**Fail**	**Proportion Pass**	**Chi-sq**
0	4965	2228	2737	0.45	
1 through 3	5295	1877	3418	0.35	
4 through 6	620	297	323	0.48	
7	119	46	73	0.39	
8 through 11	356	169	187	0.47	
12 through 17	182	70	112	0.38	$\chi^2 = 120$
18	19	9	10	0.47	7 d.f.
19+	28	7	21	0.25	$p \approx 0$
Total:	11584	4703	6881	0.41	

Figure 32. Column Chart Depicting Student Performance Based on Years of Service



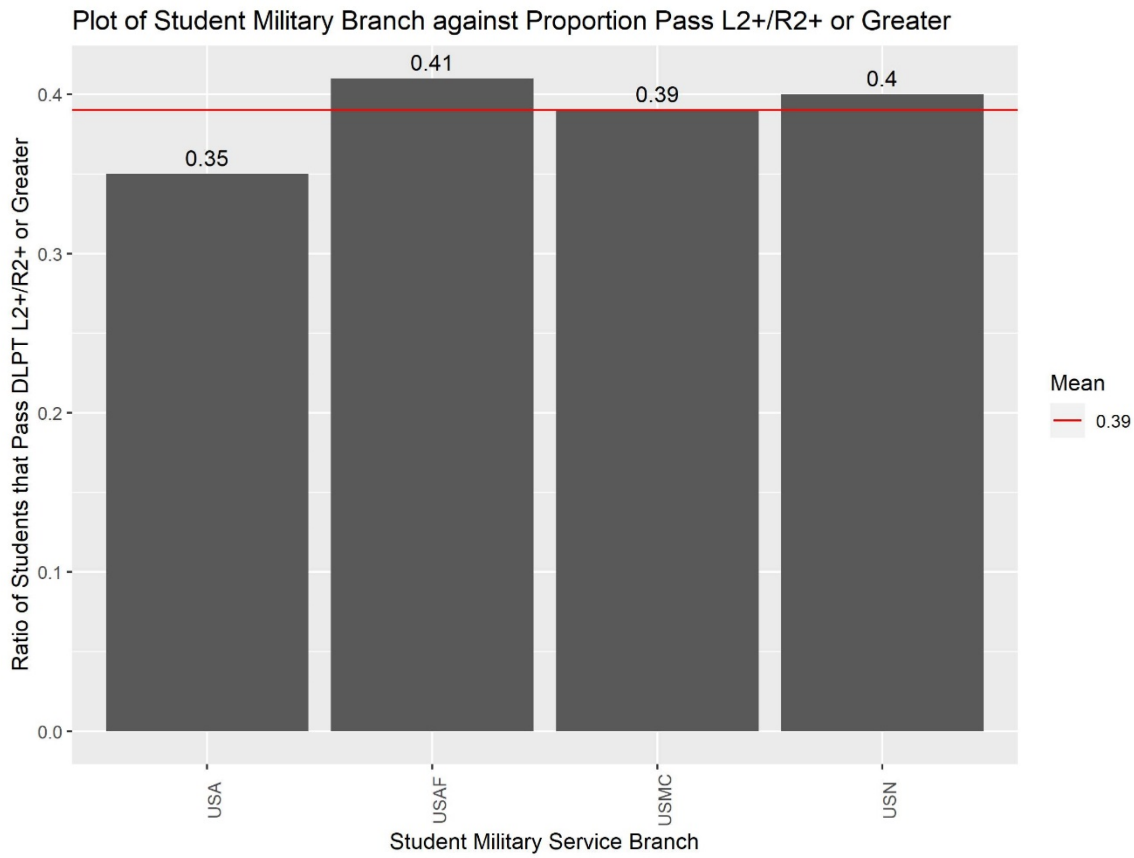
3. Results for Student's Service Branch

The results of the chi-squared goodness of fit tests for the student's service branch are shown in Table 37. A helpful visualization of the proportions of pass/fails is provided in Figure 33. We are able to reject the null hypothesis that there is no difference between the service branches for students passing the DLPT at the L2+/R2+ level or greater. This is caused by Army students' statistically poorer performance; omitting the Army results in a statistical tie. This largely confirms research that has been performed by other studies on the SDB data set going back over a decade.

Table 37. Student's Service Branch Chi-Squared Results

Student Service Branch	**Total Students**	**Pass**	**Fail**	**Proportion Pass**	**Chi-sq**
USAF	5853	2401	3452	0.41	
USN	3172	1256	1916	0.40	$\chi^2 = 36$
USMC	1913	752	1161	0.39	3 d.f.
USA	3995	1403	2592	0.35	$p \approx 0$
Total:	14933	5812	9121	0.39	

Figure 33. Column Chart Depicting Student Performance Based on Service Branch



4. Results for Student's Motivation Level

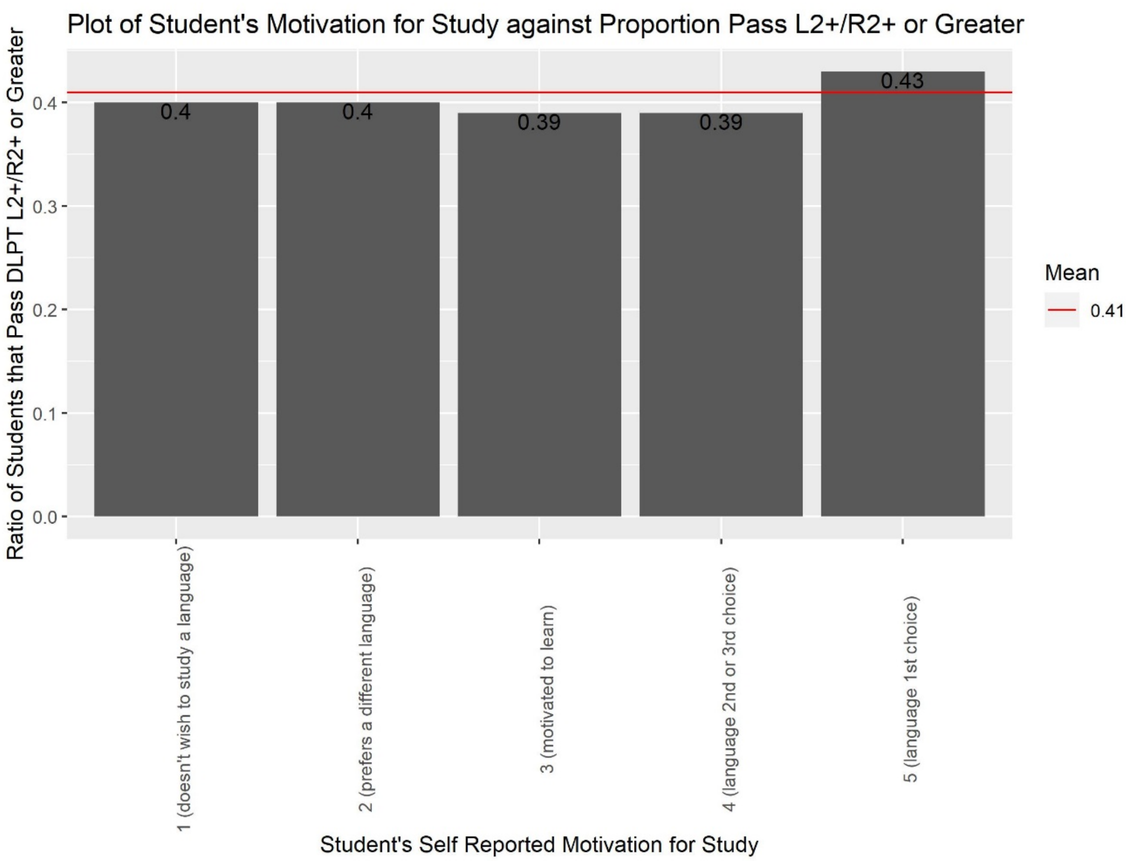
The results of the chi-squared goodness of fit tests for the student's self-reported motivation level are shown in Table 38. A helpful visualization of the proportions of pass/ fails is provided in Figure 34. Unsurprisingly the most motived students performed best.

Table 38. Student's Motivation Level Chi-Squared Results

Student Self-Reported Motivation Level	Total Students	Pass	Fail	Proportion Pass	Chi-sq
5 (motivated, language 1st choice)	4246	1832	2414	0.43	
1 (not motivated, doesn't wish to study a language)	285	115	170	0.40	

2 (not motivated, prefers a different language)	95	38	57	0.40	$\chi^2 = 18$
4 (motivated, language 2nd or 3rd choice)	2757	1085	1672	0.39	4 d.f.
3 (not preferred language, motivated to learn)	4204	1634	2570	0.39	$p \approx 0$
Total:	11587	4704	6883	0.41	

Figure 34. Column Chart Depicting Student Performance Based on Motivation



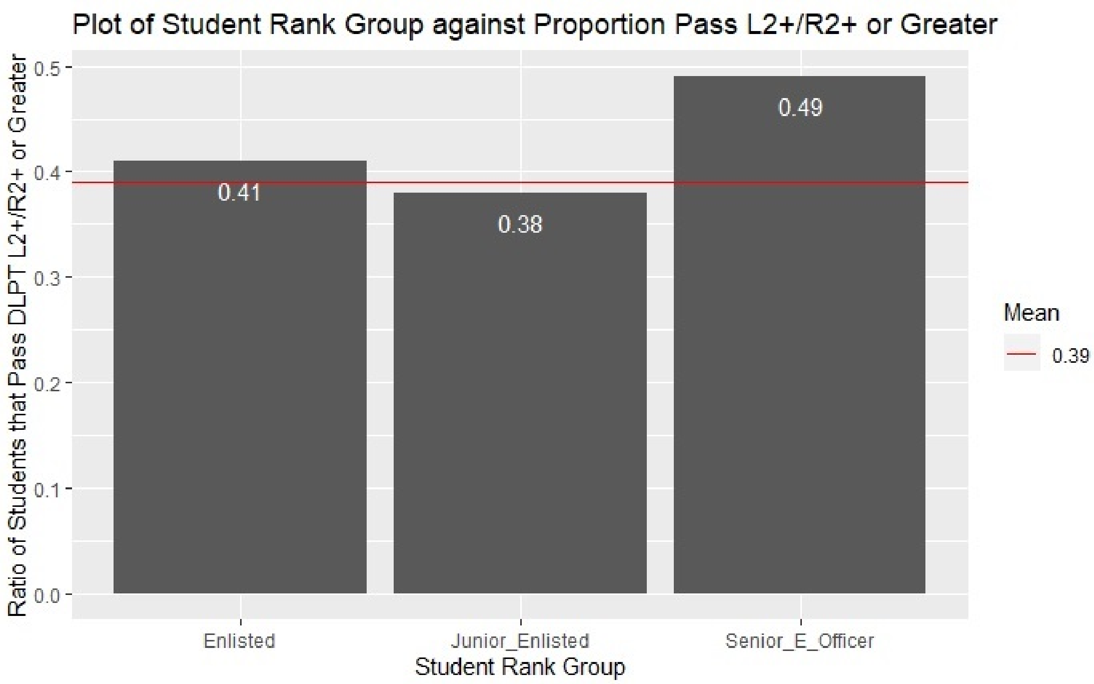
5. Results for Student Rank Group

The results of the chi-squared goodness of fit tests for the rank group are shown in Table 45. A helpful visualization of the proportions of pass/fails is provided in Figure 41. The results are not particularly surprising, because this variable is closely related to years of service and education level. The rank groups perform statistically significantly differently.

Table 39. Student Rank Group Chi-Squared Results

Student Rank Group	Total Students	Pass	Fail	Proportion Pass	Chi-sq
Senior Enlisted & Officer	529	258	271	0.49	$\chi^2 = 33$
Enlisted	3168	1301	1867	0.41	2 d.f.
Junior Enlisted	11236	4253	6983	0.38	$p \approx 0$
Total:	14933	5812	9121	0.39	

Figure 35. Column Chart Depicting Student Performance Based on Rank Group



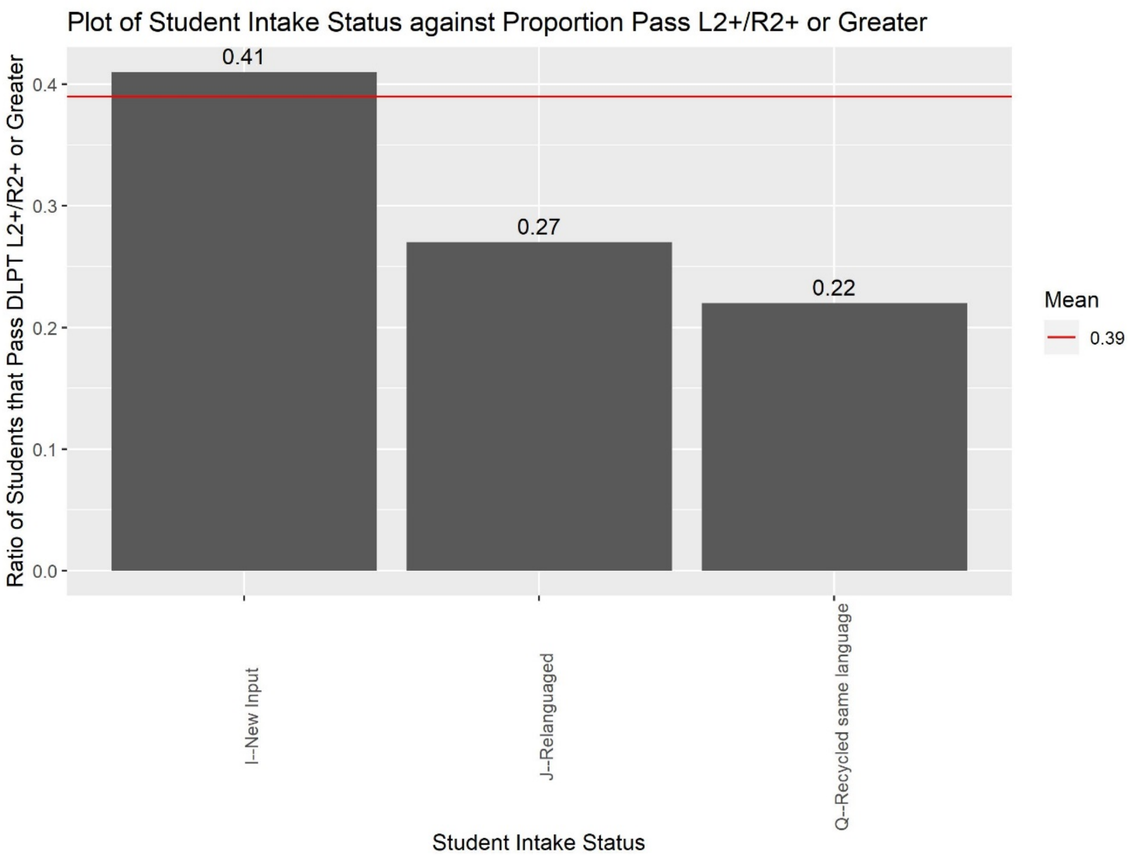
6. Results for How the Student Was Inducted into Language Training

The results of the chi-squared goodness of fit tests for student's induction status are shown in Table 40. A helpful visualization of the proportions of pass/fails is provided in Figure 36. The three statuses differ in performance. It is interesting to note that J (relanguaging) has been more successful than Q (recycling students in the same course), thought that different is more suggestive than statistically significant.

Table 40. Induction Status Chi-Squared Results

Student's In-Status	Total Students	Pass	Fail	Proportion Pass	Chi-sq
I (New Input)	13347	5456	7891	0.41	$\chi^2 = 205$
J (Relanguaged)	205	55	150	0.27	2 d.f.
Q (Recycle-Same Course)	1381	301	1080	0.22	$p \approx 0$
Total:	14933	5812	9121	0.39	

Figure 36. Column Chart for Student Performance Based on Induction Status



THIS PAGE INTENTIONALLY LEFT BLANK

V. CONCLUSIONS

A. SUMMARY

This thesis had three primary purposes. The first objective was to identify what types of machine learning models perform the best on the SDB data set. This was both to aid DLIFLC's staff in identifying students that need extra help based on an improved model's prediction, and to guide future researchers towards more fruitful analysis methods. This first purpose was accomplished with identifying random forests as the most proven modeling method on the DLIFLC SDB data set out of the three that were tried, and producing an improved model to aid DLIFLC's staff.

The second purpose of this research was to identify if a language background has significant influence on what languages a student might be successful in learning. This knowledge has the power to better inform DLIFLC's leadership about what considerations should be made when assigning students to a specific language of study. Language background, prior language source, prior language experience, and prior language proficiency are statistically significantly related to student success at the L2+/R2+ level. It might be noteworthy that students' background languages are associated with different success rates across the languages taught at DLIFLC.

The third purpose of this research was to identify any other factors or influences that exist that can be shaped into policy at DLIFLC or the recruitment process for CLAs. Anything that can be done to boost the rate of students achieving a score of L2+/R2+ or greater on the DLPT will be of great benefit. The goal of advancing student test scores without any additional resources provided to DLIFLC means that there can be no stones unturned in the data. We examined the available data in an effort to identify helpful predictors of success. Many of these other predictors like education level, years of service, and DLAB waiver are associated with student achievement on the DLPT.

B. FUTURE WORK

The DOD is shifting back towards great power competition with Russia and China and away from the Global War on Terror. These changes will bring a language realignment for languages taught at DLIFLC. These changes will bring new opportunities and more data. Many of the statistical analyses that were employed in this study suffered from not having enough data to make solid recommendations with sufficient statistical underpinning.

Many predictors showed suggestive results, but did not have enough students represented to draw concrete conclusions. These limitations may not be the case five or more years from now with continued good data collection on DLIFLC students. The statistical analysis presented in Chapter 4 of this work showed many areas with interesting differences that should be investigated in some form in the future.

The immersion program remains an area that will likely need additional study. It is still open to debate whether the program might be of much greater benefit to struggling students, and there will need to be data analysis to justify a conclusive answer to that question.

Neural networks and some of the other methods of machine learning are still in active development. Advancement in the fields of NN and data analytics might make more opportunities for improved modeling available to the SDB data set.

There are opportunities to study how students perform after graduation. Questions remain as to whether additional instruction to students who struggled on the DLPT can have a significant influence on their field performance.

The language background results of this paper can be used to develop an optimization model to select students for specific languages of study based on their language backgrounds. This would assist DLIFLC in assigning good candidates for success. Some of the predictor categories like years of accumulated service and education level show unexpected results and warrant some follow-up to see if there are some easy policy fixes that could be put into place to stave off avoidable dips in student performance.

APPENDIX A. DESCRIPTION OF VARIABLES

Table 41. Description of Variables. Adapted from Bermudez-Mendez (2020)

Name	Symbol	Classification	Description
Service Branch	Svc	Categorical	USA (Army) USN (Navy) USMC (Marine Corps) USAF (Air Force)
Language	Lang	Categorical	XX different language digraphs
Category	Lang.Cat	Categorical	Difficulty of Language: 1 (CAT I) 2 (CAT II) 3 (CAT III) 4 (CAT IV)
DLAB	DLAB	Continuous	Scores from XX to XXX
DLAB Waiver	DLAB_Waiver	Categorical	Y (Yes) N (No)
Gender	Gender	Categorical	M (Male) F (Female)
Rank	Rank	Categorical	Student Rank, E-1 to O-7
Input Status	In_Status	Categorical	I (New Input) J (Relanguaged) P (Post-DLPT) Q (Recycle – Same Course)
Output Status	Out_Status	Categorical	G (Graduate) H (Hold) L (Recycle Out Same Course) Z (Attrition)
Reason for Attrition	Reason	Categorical	* (NA) A or X (Academic Attrition)
GPA	GPA	Continuous	Scale 0.0 to 4.0
DLPT Listening	DLPT.L	Categorical	00 (L0) 06 (L0+) 10 (L1) 16 (L1+) 20 (L2) 26 (L2+) 30 (L3)

Name	Symbol	Classification	Description
DLPT Reading	DLPT.R	Categorical	00 (R0) 06 (R0+) 10 (R1) 16 (R1+) 20 (R2) 26 (R2+) 30 (R3)
DLPT OPI	OPI.S	Categorical	06 (L0+) 10 (L1) 16 (L1+) 20 (L2) 26 (L2+) 30 (L3)
Years of Service	Yrs.Svc	Numerical	Range 1 - 41
Marital Status	Marital.St	Categorical	*Blank* (No input from Student) S (Single) M (Married)
Education Level	Edu	Categorical	1 (Non-High School) 2 (High School or GED) 3 (1 Year College) 4 (2 Years College) 5 (3 Years College) 6 (4 Years College) 7 (Bachelor's Degree) 8 (Master's Degree) 9 (Doctorate) 0 or NA (No input from Student)
Motivation	Motive	Categorical	1 (Not Motivated, does not want to study any language) 2 (Not Motivated, prefers another language) 3 (Not My preferred language, but motivated to learn) 4 (Motivated, language is second or third choice) 5 (Motivated, language is first choice)
Prior Language	Prior.Lang	Categorical	130 various languages
Native English Speaker	Native.Eng	Categorical	*Blank* (No Student Response)

Name	Symbol	Classification	Description
			Y (Yes) N (No)
Native Other Speaker	Native.Oth	Categorical	*Blank* (No Student Response) Y (Yes) N (No)
Birthdate	Birthdate	Numeric	Age Range xx – xx
Prior Lang. Proficiency	Prior.Lang.Prof	Categorical	A (Poor) B (Fair) C (Good) D (Excellent) X (None)
Prior Lang. Source	Prior.Lang.Src	Categorical	A (Civilian School) B (DLI) C (Foreign Residence) D (Home Environment) E (Military – Other than DLI) F (Self Study) X (NA)
Prior Lang. Experience	Prior.Lang.Exp	Categorical	A (Conversation) B (Instructor) C (Interpreter) D (Interrogator) E (Reader) F (Transcriber) G (Translator) X (None)
Language Immersion	Immersion	Categorical	O (OCONUS Immersion) C (CONUS Immersion) U (Immersion location not provided) *Blank* (No Immersion)
Elementary Lang I	FL101	Categorical	Student's Letter Grade
Elementary Lang II	FL102	Categorical	Student's Letter Grade
Elementary Convo.	FL110	Categorical	Student's Letter Grade
Intro to Job Related Skills in Lang.	FL120	Categorical	Student's Letter Grade
Intro to Lang Culture	FL140	Categorical	Student's Letter Grade
Intermediate Lang I	FL201	Categorical	Student's Letter Grade
Intermediate Lang II	FL202	Categorical	Student's Letter Grade
Intermediate Convo.	FL210	Categorical	Student's Letter Grade

Name	Symbol	Classification	Description
Intro to Military Topics in Lang.	FL220	Categorical	Student's Letter Grade
History and Geography of Lang Region	FL240	Categorical	Student's Letter Grade
Advanced Lang I	FL301	Categorical	Student's Letter Grade
Advanced Lang II	FL302	Categorical	Student's Letter Grade
Advanced Convo.	FL310	Categorical	Student's Letter Grade
Comprehensive Military Topics in Lang	FL320	Categorical	Student's Letter Grade
Lang Area/Cultural Studies	FL340	Categorical	Student's Letter Grade

APPENDIX B. LETTER-GRADE-TO-GPA CONVERSION TABLE

Table 42. Letter-Grade-To-GPA Conversion Table (Bermudez-Mendez 2020)

Letter Grade	GPA
A	4.0
A-	3.7
B+	3.3
B	3.0
B-	2.7
C+	2.3
C	2.0
C-	1.7
D+	1.3
D	1.0
F	0
P	2.0

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX C. LANGUAGES TAUGHT AT DLIFLC

Table 43. Languages Taught at DLIFLC (Bermudez-Mendez 2020)

Language	Digraph	Category
Pashtu-Afghan	PV	4
Chinese-Mandarin	CM	4
Japanese	JA	4
Arabic-Iraqi	DG	4
Arabic-Egyptian	AE	4
Arabic-Levantine Syrian	AP	4
Korean	KP	4
Arabic-Modern Standard	AD	4
Russian	RU	3
Urdu	UR	3
Tagalong	TA	3
Hebrew-Modern	HE	3
Persian-Farsi	PF	3
Indonesian	JN	2
Spanish	QB	1
French	FR	1

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX D. PRIOR LANGUAGES OBSERVED AT DLIFLC AND GROUPING

Table 44. Prior Languages Observed at DLIFLC and Grouping

Original Digraph Code	Language	Number of Students	Language Grouping	Total in Language Grouping
XX	None	3349	None	3349
XO	Other	2978	Other	
SP	Sotho	11	Other	
TL	Tigrinya	12	Other	
AC	Amharic	2	Other	
JB	IBO	1	Other	
LS	Luganda	1	Other	
QT	Otetela	2	Other	
QU	Quechua	1	Other	
RN	Kirundi	1	Other	
ZZ	Unknown	93	Other	3102
AN	Arabic-Saudi	1	Arabic	
AV	Arabic-Sudanese	1	Arabic	
QE	Arabic-Modern Standard	1	Arabic	
AE	Arabic-Egyptian	12	Arabic	
AP	Arabic-Levantine Syrian	12	Arabic	
AD	Arabic-Modern Standard	280	Arabic	
DG	Arabic-Gulf	12	Arabic	318
PF	Persian-Iranian	53	Indo-Iranian	
PG	Persian-Afghan aka Dari	19	Indo-Iranian	
UX	Uzbek	2	Indo-Iranian	
QP	Punjabi-Western	12	Indo-Iranian	
HJ	Hindi	8	Indo-Iranian	
QA	Oriya	1	Indo-Iranian	
PJ	Punjabi	1	Indo-Iranian	
KA	Kanarese	1	Indo-Iranian	
GQ	Gondi	1	Indo-Iranian	
UR	Urdu	12	Indo-Iranian	
PV	Pashtu-Afghan aka Pashto	19	Indo-Iranian	
XS	Kurdish-Central	1	Indo-Iranian	
KU	Kurdish	1	Indo-Iranian	131
GM	German	817	German	
SY	Swedish	12	German	
DA	Danish	4	German	
DU	Dutch	4	German	
GN	Scottish-Gaelic	4	German	

Original Digraph Code	Language	Number of Students	Language Grouping	Total in Language Grouping
AA	Afrikaans	3	German	
NR	Norwegian	3	German	
FJ	Finnish	4	German	
GF	Irish	1	German	852
FR	French	1119	French	
HC	Hattian-Creole	2	French	1121
LA	Spanish-American	8	Spanish	
QB	Spanish	4303	Spanish	
QC	Spanish-Caribbean	2	Spanish	4313
JS	Italian-Sicilian	320	Romance	
JT	Italian	108	Romance	
CB	Catalan	1	Romance	
LH	Latin-Ecclesiastic	1	Romance	
VL	Latin	11	Romance	
YL	Latin	2	Romance	
RQ	Romanian	12	Romance	
AB	Albanian	2	Romance	
PQ	Portuguese-Brazilian	44	Romance	
PT	Portuguese-European	6	Romance	
PY	Portuguese	21	Romance	
PR	Provencal	1	Romance	529
RU	Russian	237	Balto-Slavic	
UK	Ukrainian	1	Balto-Slavic	
SK	Slovak	3	Balto-Slavic	
SC	Serbo-Croatian	44	Balto-Slavic	
BU	Bulgarian	2	Balto-Slavic	
PL	Polish	3	Balto-Slavic	
KR	Kashubian	2	Balto-Slavic	
LT	Lithuanian	3	Balto-Slavic	295
GE	Greek-New Testament	1	Greek-Oghuz	
GR	Greek	34	Greek-Oghuz	
TU	Turkish	16	Greek-Oghuz	
AX	Azerbaijani	1	Greek-Oghuz	
HU	Hungarian	5	Greek-Oghuz	
AR	Armenian	1	Greek-Oghuz	58
KP	Korean	206	Korean	206
JA	Japanese	159	Japanese	159
CA	Cambodian	2	Austroasiatic	
TH	Thai	13	Austroasiatic	
VN	Vietnamese-Hanoi	25	Austroasiatic	40
QC	Malagasy	2	Austronesian	
TA	Tagalong	90	Austronesian	

Original Digraph Code	Language	Number of Students	Language Grouping	Total in Language Grouping
JN	Indonesian	8	Austronesian	
LC	Lao	1	Austronesian	
LY	Murano	1	Austronesian	
MD	Madurese	1	Austronesian	
PD	Palauan	1	Austronesian	
SA	Samoan	1	Austronesian	
MG	Malagasy	2	Austronesian	
QR	Kusaie	2	Austronesian	
WW	Waray-Waray	2	Austronesian	
QV	Pampangan	4	Austronesian	115
HE	Hebrew	39	Hebrew	39
CH	Chinese-Hakka	1	Chinese	
CN	Chinese-Anhwei	1	Chinese	
CM	Chinese-Mandarin	287	Chinese	
CC	Chinese-Cantonese	16	Chinese	
CH	Chinese-Hakka	1	Chinese	306

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX E. BERMUDEZ-MENDEZ’S 3RD SEMESTER LOGISTIC REGRESSION MODEL OUTPUT

Figure 37. AUC for Bermudez-Mendez’s 3rd Semester Logistic Regression Model

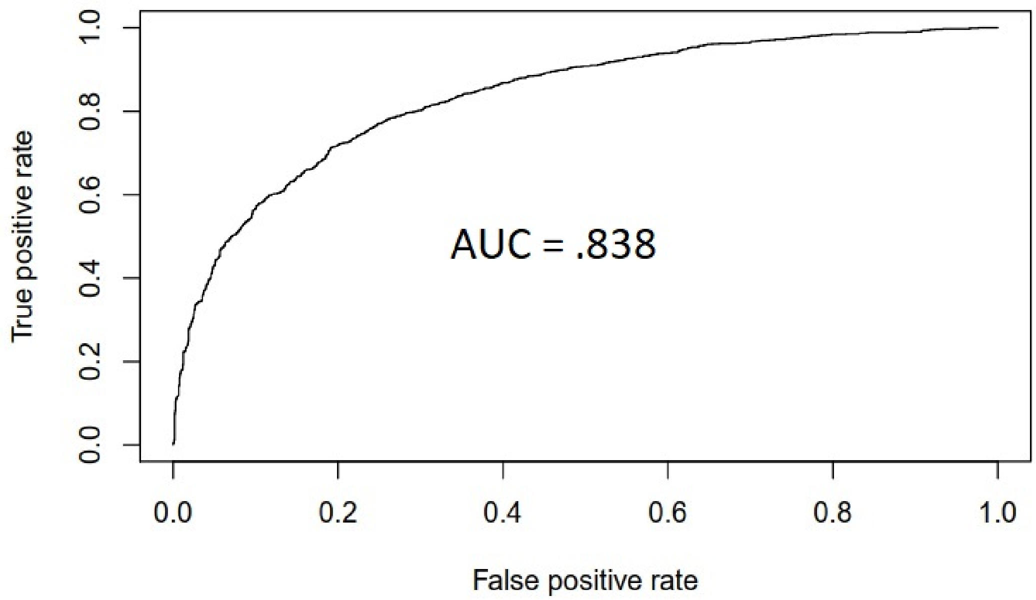


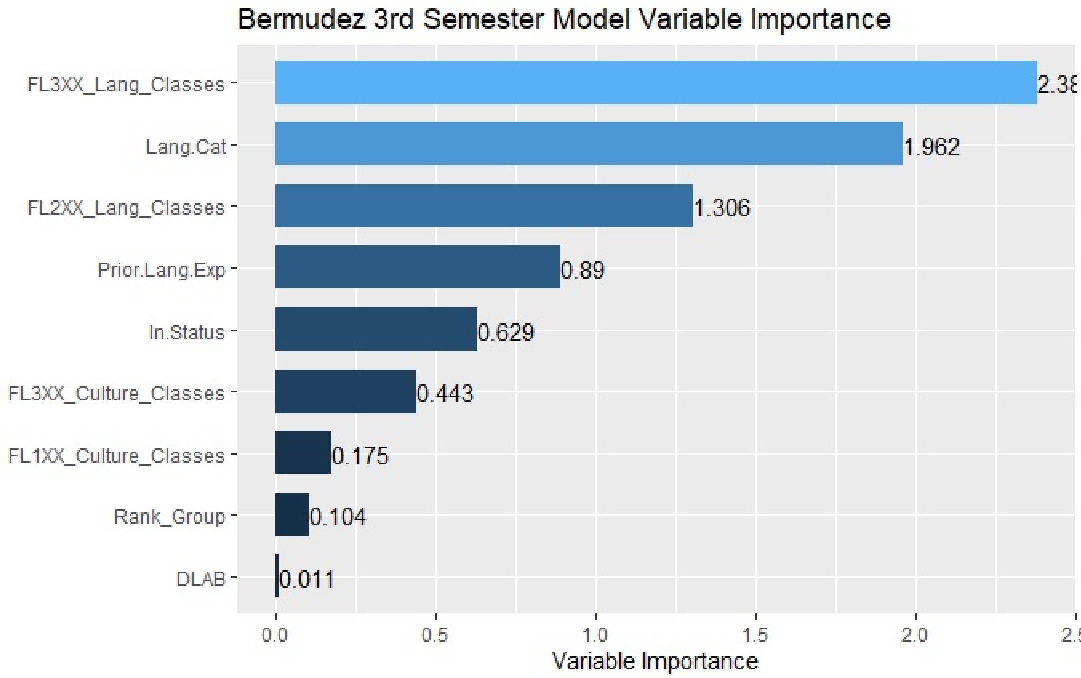
Table 45. Bermudez-Mendez’s 3rd Semester Logistic Regression Model Classification Table

Predicted/Observed	**Success**	**Failure**	
Success	**1481**	**366**	1847
Failure	**333**	**800**	1133
	1814	1166	

Table 46. Bermudez-Mendez’s 3rd Semester Logistic Regression Model Classification Derived Metrics

Sensitivity	0.80
Specificity	0.71
Precision (PPV)	0.82
(NPV)	0.69
Overall Accuracy	0.77
AUC:	0.838

Figure 38. Bermudez-Mendez's 3rd Semester Logistic Regression Variable Importance



APPENDIX F. RESULTS FROM LANGUAGE BACKGROUND STUDY

The rest of the results presented here will be from the chi-squared test statistics shown in tabulated form. The total number of students for each language will be highlighted in red if the predicted numbers that did not meet the minimum criterion for having at least five students. The results in each table will be sorted on their proportion of student's success from high to low. The proportion is simply the number of students who achieved L2+/R2+ or greater on the DLPT divided by total students for that particular category. The larger the proportion the better the students did. There is a line drawn across the tables indicating where the cutoff for the mean proportion of students for that language occurred. Each language was tested in a two-way chi-square test against the students with no language background. The two-way table for the one instance of students with no language background is tested against the sum results of all the students with a language background.

Table 47. Tabulated Results for Language Background Study for Combined Arabic (AD, AE, AP, DG) Students

Language Background Group	**Total Students Taking Arabic**	**Pass**	**Fail**	**Proportion Pass**	**Chi-sq *p*-value**
KP-Korean	41	19	22	0.46	< .05
AE-Austronesian	19	8	11	0.42	> .05
FR-French	274	112	162	0.41	< .01
XO-Other	70	28	42	0.40	> .05
HE-Hebrew	13	5	8	0.38	> .05
AD-Arabic	97	37	60	0.38	> .05
GM-German	218	83	135	0.38	< .05
BS-Balto-Slavic	45	17	28	0.38	> .05
GO-Greek-Oghuz	8	3	5	0.38	> .05
CM-Chinese	46	16	30	0.35	> .05
AA-Austroasiatic	6	2	4	0.33	> .05
LA-Spanish	1021	337	684	0.33	> .05
PF-Indo-Iranian	19	6	13	0.32	> .05
JA-Japanese	16	5	11	0.31	> .05

Language Background Group	Total Students Taking Arabic	Pass	Fail	Proportion Pass	Chi-sq *p*-value
XX-None-English	1972	594	1378	0.30	< .01
RM-Romance	130	34	96	0.26	> .05
Total:	3995	1306	2689	0.33	

Table 48. Tabulated Results for Language Background Study for Chinese-Mandarin Students

Language Background Group	Total Students Taking Chinese-Mandarin	Pass	Fail	Proportion Pass	Chi-sq *p*-value
CM-Chinese	106	69	37	0.65	p < .05
KP-Korean	47	30	17	0.64	p > .05
AE-Austronesian	19	12	7	0.63	p > .05
AD-Arabic	23	14	9	0.61	p > .05
RM-Romance	157	93	64	0.59	p > .05
JA-Japanese	31	18	13	0.58	p > .05
FR-French	186	107	79	0.58	p > .05
AA-Austroasiatic	11	6	5	0.55	p > .05
LA-Spanish	716	388	328	0.54	p > .05
GM-German	136	73	63	0.54	p > .05
XX-None-English	483	250	233	0.52	p > .05
PF-Indo-Iranian	8	4	4	0.50	p > .05
GO-Greek-Oghuz	10	5	5	0.50	p > .05
BS-Balto-Slavic	37	16	21	0.43	p > .05
XO-Other	37	15	22	0.41	p > .05
HE-Hebrew	3	1	2	0.33	p > .05
Total:	2010	1101	909	0.55	

Table 49. Tabulated Results for Language Background Study for French Students

Language Background Group	**Total Students Taking French**	**Pass**	**Fail**	**Proportion Pass**	**Chi-sq *p*-value**
HE-Hebrew	2	2	0	1.00	> .05
AD-Arabic	90	62	28	0.69	~ 0
GO-Greek-Oghuz	3	2	1	0.67	> .05
FR-French	49	28	21	0.57	< .01
AE-Austronesian	2	1	1	0.50	> .05
GM-German	20	10	10	0.50	< .05
RM-Romance	15	7	8	0.47	> .05
LA-Spanish	110	35	75	0.32	> .05
BS-Balto-Slavic	13	4	9	0.31	> .05
XO-Other	7	2	5	0.29	NA
XX-None-English	166	40	126	0.24	~ 0
KP-Korean	11	2	9	0.18	> .05
PF-Indo-Iranian	14	1	13	0.07	> .05
CM-Chinese	4	0	4	0.00	> .05
JA-Japanese	3	0	3	0.00	> .05
AA-Austroasiatic	0	0	0	infinite	NA
Total:	509	196	313	0.39	

Table 50. Tabulated Results for Language Background Study for Hebrew Students

Language Background Group	Total Students taking Hebrew	Pass	Fail	Proportion Pass	Chi-sq *p*-value
GM-German	20	20	0	1.00	> .05
GO-Greek-Oghuz	5	5	0	1.00	> .05
KP-Korean	4	4	0	1.00	> .05
HE-Hebrew	4	4	0	1.00	> .05
AD-Arabic	2	2	0	1.00	> .05
RM-Romance	5	4	1	0.80	> .05
JA-Japanese	5	4	1	0.80	> .05
XX-None-English	95	74	21	0.78	> .05
LA-Spanish	104	81	23	0.78	> .05
BS-Balto-Slavic	8	6	2	0.75	> .05
PF-Indo-Iranian	3	2	1	0.67	> .05
FR-French	21	14	7	0.67	> .05
XO-Other	6	4	2	0.67	> .05
CM-Chinese	5	3	2	0.60	> .05
AE-Austronesian	2	0	2	0.00	> .05
AA-Austroasiatic	0	0	0	Infinite	NA
Total:	289	227	62	0.79	

Table 51. Tabulated Results for Language Background Study for Indonesian Students

Language Background Group	Total Students Taking Indonesian	Pass	Fail	Proportion Pass	Chi-sq *p*-value
BS-Balto-Slavic	3	3	0	1.00	p > .05
JA-Japanese	2	2	0	1.00	p > .05
AA-Austroasiatic	1	1	0	1.00	p > .05
PF-Indo-Iranian	1	1	0	1.00	p > .05
RM-Romance	1	1	0	1.00	p > .05
CM-Chinese	20	17	3	0.85	p > .05
LA-Spanish	27	24	3	0.89	p < .05
KP-Korean	7	6	1	0.86	p > .05
AE-Austronesian	6	5	1	0.83	p > .05
GM-German	5	4	1	0.80	p > .05
AD-Arabic	4	3	1	0.75	p > .05
FR-French	6	4	2	0.67	p > .05
XO-Other	3	2	1	0.67	NA
XX-None-English	76	49	27	0.64	p < .01
GO-Greek-Oghuz	0	0	0	Infinite	NA
HE-Hebrew	0	0	0	Infinite	NA
Total:	162	122	40	0.75	

Table 52. Tabulated Results for Language Background Study for Korean Students

Language Background Group	Total Students Taking Korean	Pass	Fail	Proportion Pass	Chi-sq *p*-value
HE-Hebrew	1	1	0	1.00	> .05
KP-Korean	31	22	9	0.71	~ 0
AD-Arabic	6	4	2	0.67	> .05
PF-Indo-Iranian	2	1	1	0.50	> .05
RM-Romance	52	23	29	0.44	< .05
CM-Chinese	26	10	16	0.38	> .05
LA-Spanish	303	108	195	0.36	> .05
FR-French	88	31	57	0.35	> .05
XO-Other	12	4	8	0.33	> .05
JA-Japanese	53	17	36	0.32	> .05
XX-None-English	767	225	542	0.29	< .01
GM-German	59	17	42	0.29	> .05
BS-Balto-Slavic	15	4	11	0.27	> .05
AE-Austronesian	8	2	6	0.25	> .05
GO-Greek-Oghuz	8	2	6	0.25	> .05
AA-Austroasiatic	7	1	6	0.14	> .05
Total:	1437	471	966	0.33	

Table 53. Tabulated Results for Language Background Study for Persian-Farsi Students

Language Background Group	Total Students Taking Persian-Farsi	Pass	Fail	Proportion Pass	Chi-sq *p*-value
AA-Austroasiatic	2	2	0	1.00	> .05
PF-Indo-Iranian	41	23	18	0.56	< .05
CM-Chinese	23	12	11	0.52	> .05
AE-Austronesian	21	9	12	0.43	> .05
AD-Arabic	26	11	15	0.42	> .05
KP-Korean	24	10	14	0.42	> .05
HE-Hebrew	5	2	3	0.40	> .05
GM-German	91	36	55	0.40	> .05
BS-Balto-Slavic	35	13	22	0.37	> .05
XX-None-English	1154	412	742	0.36	> .05
LA-Spanish	496	171	325	0.34	> .05
FR-French	149	47	102	0.32	> .05
XO-Other	26	7	19	0.27	> .05
RM-Romance	45	10	35	0.22	> .05
GO-Greek-Oghuz	3	0	3	0.00	> .05
JA-Japanese	1	0	1	0.00	> .05
Total:	2142	765	1377	0.36	

Table 54. Tabulated Results for Language Background Study for Pashto (Pashtu-Afghan) Students

Language Background Group	**Total Students Taking Pashto**	**Pass**	**Fail**	**Proportion Pass**	**Chi-sq *p*-value**
GO-Greek-Oghuz	6	6	0	1.00	< .01
JA-Japanese	2	2	0	1.00	> .05
PF-Indo-Iranian	10	9	1	0.90	< .01
AD-Arabic	17	13	4	0.76	< .01
AA-Austroasiatic	3	2	1	0.67	> .05
KP-Korean	9	6	3	0.67	> .05
FR-French	74	48	26	0.65	~ 0
CM-Chinese	22	14	8	0.64	< .05
AE-Austronesian	8	5	3	0.63	> .05
RM-Romance	41	25	16	0.61	< .01
HE-Hebrew	5	3	2	0.60	> .05
LA-Spanish	281	168	113	0.60	~ 0
BS-Balto-Slavic	22	13	9	0.59	> .05
GM-German	69	39	30	0.57	< .01
XO-Other	416	233	183	0.56	< .01
XX-None-English	80	28	52	0.35	~ 0
Total:	1065	614	451	0.58	

Table 55. Tabulated Results for Language Background Study for Spanish Students

Language Background Group	**Total Students Taking Spanish**	**Pass**	**Fail**	**Proportion Pass**	**Chi-sq *p*-value**
GO-Greek-Oghuz	7	5	2	0.71	< .05
AE-Austronesian	13	9	4	0.69	< .01
HE-Hebrew	2	1	1	0.50	> .05
RM-Romance	34	13	21	0.38	> .05
KP-Korean	11	4	7	0.36	> .05
AA-Austroasiatic	6	2	4	0.33	> .05
AD-Arabic	29	9	20	0.31	> .05
FR-French	112	34	78	0.30	> .05
GM-German	89	27	62	0.30	> .05
XX-None-English	414	104	310	0.25	> .05
BS-Balto-Slavic	24	6	18	0.25	> .05
LA-Spanish	658	162	496	0.25	> .05
JA-Japanese	21	4	17	0.19	> .05
CM-Chinese	23	4	19	0.17	> .05
XO-Other	28	4	24	0.14	> .05
PF-Indo-Iranian	17	1	16	0.06	> .05
Total:	1488	389	1099	0.26	

Table 56. Tabulated Results for Language Background Study for Russian Students

Language Background Group	Total Students Taking Russian	Pass	Fail	Proportion Pass	Chi-sq *p*-value
HE-Hebrew	1	1	0	1.00	> .05
PF-Indo-Iranian	7	4	3	0.57	> .05
JA-Japanese	23	13	10	0.57	< .05
CM-Chinese	22	11	11	0.50	< .05
KP-Korean	4	2	2	0.50	> .05
BS-Balto-Slavic	77	37	40	0.48	< .01
AD-Arabic	10	4	6	0.40	> .05
GO-Greek-Oghuz	5	2	3	0.40	> .05
RM-Romance	24	9	15	0.38	> .05
FR-French	124	43	81	0.35	> .05
GM-German	118	39	79	0.33	> .05
XO-Other	31	10	21	0.32	> .05
AE-Austronesian	10	3	7	0.30	> .05
LA-Spanish	440	130	310	0.30	> .05
XX-None-English	398	109	289	0.27	< .05
AA-Austroasiatic	0	0	0	Infinite	NA
Total:	1294	417	877	0.48	

Table 57. Tabulated Results for Language Background Study for Tagalong Students

Language Background Group	**Total Students Taking Tagalog**	**Pass**	**Fail**	**Proportion Pass**	**Chi-sq *p*-value**
AA-Austroasiatic	4	4	0	1.00	> .05
XO-Other	3	3	0	1.00	> .05
HE-Hebrew	1	1	0	1.00	> .05
GO-Greek-Oghuz	1	1	0	1.00	> .05
JA-Japanese	1	1	0	1.00	> .05
KP-Korean	9	7	2	0.78	> .05
CM-Chinese	4	3	1	0.75	> .05
AE-Austronesian	4	3	1	0.75	> .05
GM-German	12	9	3	0.75	> .05
BS-Balto-Slavic	3	2	1	0.67	> .05
XX-None-English	40	26	14	0.65	> .05
FR-French	13	8	5	0.62	> .05
LA-Spanish	62	36	26	0.58	> .05
RM-Romance	7	3	4	0.43	> .05
AD-Arabic	0	0	0	Infinite	NA
PF-Indo-Iranian	0	0	0	Infinite	NA
Total:	164	107	57	0.65	

Table 58. Tabulated Results for Language Background Study for Urdu Students

Language Background Group	Total Students Taking Urdu	Pass	Fail	Proportion Pass	Chi-sq *p*-value
AE-Austronesian	2	2	0	1.00	> .05
AD-Arabic	11	6	5	0.55	< .05
CM-Chinese	4	2	2	0.50	> .05
GM-German	11	5	6	0.45	> .05
BS-Balto-Slavic	10	4	6	0.40	> .05
XO-Other	77	16	61	0.21	> .05
FR-French	19	6	13	0.32	> .05
PF-Indo-Iranian	8	2	6	0.25	> .05
RM-Romance	12	3	9	0.25	> .05
LA-Spanish	85	19	66	0.22	> .05
XX-None-English	73	14	59	0.19	> .05
KP-Korean	6	1	5	0.17	> .05
HE-Hebrew	2	0	2	0.00	> .05
GO-Greek-Oghuz	2	0	2	0.00	> .05
JA-Japanese	1	0	1	0.00	> .05
AA-Austroasiatic	0	0	0	Infinite	NA
Total:	323.00	80	243	0.25	

LIST OF REFERENCES

Ancient Origins (2014) New study suggests that the Philippines is the ancestral homeland of Polynesians. Accessed January 20, 2021, https://www.ancient-origins.net/news-evolution-human-origins/new-study-suggests-philippines-ancestral-homeland-polynesians-001463.

Anderson RE (1997) Study of initial entry student attrition from Defense Language Institute Foreign Language Center. Master's thesis, Naval Postgraduate School, Monterey, CA, http://hdl.handle.net/10945/9027.

Augustyn A (2018) Hebrew language. Encyclopedia Britannica. Accessed January 13, 2021, https://www.britannica.com/topic/Hebrew-language.

Bermudez-Mendez (2020) Student Success Factors At Defense Language Institute Foreign Language Center. Master's Thesis, Operations Research Department, Naval Postgraduate School, Monterey, CA, https://calhoun.nps.edu/handle/10945/64866.

Bhandari A (2020) AUC-ROC curve in machine learning clearly explained. Analytics Vidhya. Accessed January 12, 2021, https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/.

Blust RA (2018) Austronesian languages. Encyclopedia Britannica. Accessed January 20, 2021, https://www.britannica.com/topic/Austronesian-languages.

Bottou L (2012) Stochastic gradient descent tricks. Microsoft Research, Redmond, WA. Accessed January 20, 2021, https://cilvr.cs.nyu.edu/diglib/lsml/bottou-sgd-tricks-2012.pdf

Breiman L, Friedman J, Olshen R, Stone C (1984) *Classification and Regression Trees.* (Wadsworth International Group, Belmont, CA).

Breiman L (2001) Random forests. *Mach Learn* 45(1):5–3, https://doi.org/10.1023/A:1010933404324.

Buccini AF (1998) Germanic language. Encyclopedia Britannica. Accessed January 20, 2021, https://www.britannica.com/topic/Germanic-languages.

Cardona G (1998) Indo-Iranian languages. Encyclopedia Britannica. Accessed January 20, 2021, https://www.britannica.com/topic/Indo-Iranian-languages.

Cook D (2017) Practical Machine Learning with H2O: P-valueful, Scalable Techniques for AI and Deep Learning (O'Reilly, Sebastopol, CA).

Defense Language and National Security Education Office (2020) Department of Defense language codes list. Accessed January 13, 2021, https://dlnseo.org/sites/default/files/2020%20DoD%20Language%20Codes%20List%201-31-2020.pdf.

Defense Language Institute Foreign Language Center (2021) Mission vision and values. Accessed January 5, 2021, https://www.dliflc.edu/about/mission-vision/.

Defense Language Institute Foreign Language Center (2020) General catalog 2019–2020, v10c (Monterey, CA). https://www.dliflc.edu/wp-content/uploads/2018/11/DLIFLC_catalog_2019-20_v10c.pdf.

Defense Language Institute Foreign Language Center (2017) Institutional self-evaluation. Report, DLIFLC, Monterey, CA. https://www.dliflc.edu/wpcontent/uploads/2018/01/DLIFLC-Self-Study-December-2017_small.pdf.

Department of the Army (2015) Defense Language Institute Foreign Language Center (DLIFLC) plan to achieve 2+/2+ executive summary. Memorandum, Washington, DC.

Department of Defense (2009) DOD language testing program. DOD Instruction 5160.71, Washington, DC, https://dlnseo.org/sites/default/files/DoDI_5160.71.pdf.

Devore JL (2016) *Probability and Statistics for Engineering and the Sciences.* (Cengage Learning, Boston, MA).

Diffloth G (2018) Austroasiatic languages. Encyclopedia Britannica. Accessed 20 January 2021, https://www.britannica.com/topic/Austroasiatic-languages.

DLABprep.com (2010) DLAB test score range. Accessed January 6, 2021, https://web.archive.org/web/20160425025151/http://dlabprep.com/dlab-test-score-range/.

Goodfellow I, Bengio Y, Courville A (2016) *Deep Learning*. http://www.deeplearningbook.org.

H2o.ai (2021) Deep Learning (Neural Networks), algorithms, docs. Accessed January 8, 2020. http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/deep-learning.html.

Haykin S (2009) *Neural Networks and Learning Machines* (Prentice Hall, Upper Saddle River, NJ).

Haupt AC (2014) Analysis of Korean academic attrition at the Defense Language Institute Foreign Language Center. Master's thesis, Operations Research Department, Naval Postgraduate School, Monterey, CA.

Hosmer DW, Lemeshow S, Sturdivant RX (2013) Applied Logistic Regression Third Edition (John Wiley & Sons, Hoboken, NJ).

Indo-Iranian Languages (2021) *Wikipedia*. Accessed January 20, 2021, https://en.wikipedia.org/wiki/Indo-Iranian_languages.

Kim J, Bassett R, Whitaker L (2020) Automating vessel detection with passive sonar signals and convolutional neural networks. Master's thesis, Operations Research Department, Naval Postgraduate School, Monterey, CA. http://hdl.handle.net/10945/66091.

LeDell E (2020) R Interface for the 'h2o' Scalable Machine Learning Platform. https://cran.r-project.org/web/packages/h2o/h2o.pdf.

Li FF, Johnson J, Yeung S (2020) Convolutional neural networks: architectures, convolution / pooling layers. Course notes, CS231n: Convolutional Neural Networks for Visual Recognition, Spring Quarter, Department of Computer Science, Stanford University, Stanford, CA. https://cs231n.github.io/convolutional-networks/.

Li Q (2014) Research on age-related factors in foreign language learning. College of Foreign Languages, China Three Gorges University, China. International Journal of Language and Linguistics. Vol. 2, No. 1, 2014, pp. 31–37 http://www.sciencepublishinggroup.com/journal/paperinfo?journalid=501&doi=10.11648/j.ijll.20140201.14.

Martin SE (2019). Korean language. Encyclopedia Britannica. Accessed January 20, 2021, https://www.britannica.com/topic/Korean-language.

Military.com (2020a) What's a cryptologic linguist? Accessed January 5, 2021, https://www.military.com/join-armed-forces/army-cryptologic-linguist.html

Military.com (2020b) ASVAB scores and Air Force jobs. Accessed January 5, 2021, https://www.military.com/join-armed-forces/asvab/asvab-and-air-force-jobs.html.

Military.com (2020c) ASVAB scores and Navy jobs. Accessed January 5, 2021, https://www.military.com/join-armed-forces/asvab-and-navy-mos-jobs.html.

Oghuz Languages (2021) Wikipedia. Accessed January 20, 2021, https://en.wikipedia.org/wiki/Oghuz_languages.

Personnel Testing Division. (2009). ASVAB technical bulletin No. 4: P&P-ASVAB Forms 23–27. Defense Manp-value Data Center.

Pham CK (2019) Predicting the next port visit of a vessel using AIS data. Master's thesis, Naval Postgraduate School, Monterey, CA, http://hdl.handle.net/10945/64046.

R Core Team (2020) R: A language and environment for statistical computing. R Foundation for Statistical Computing, http://www.R-project.org.

Ruijgh CJ, Newton BE Malikouti-Drachman A and Lejeune M (2018) Greek language. Encyclopedia Britannica. Accessed January 20, 2021, https://www.britannica.com/topic/Greek-language.

Sala M (1999) Romance languages. Encyclopedia Britannica. Accessed January 20, 2020, https://www.britannica.com/topic/Romance-languages.

Schmitz EJ, Stoloff PH, Wolfanger JS, & Sayala S (2009). Accession Screening for Language Skills and Abilities. Alexandria: CNA.

Shaw V M W and Lett J A (1993) Relationships of language aptitude and age to DLPT results among senior officer students in DLIFLC Basic Language Courses.

Shibatani M (2019). Japanese language. Encyclopedia Britannica. Accessed January 20, 2021, https://www.britannica.com/topic/Japanese-language.

Sirsat M (2019) What is a confusion matrix and advanced classification metrics. Accessed January 12 2021, https://manisha-sirsat.blogspot.com/2019/04/confusionmatrix.html.

Encyclopedia Britannica (1998) Balto-Slavic languages. Accessed January 20, 2021, https://www.britannica.com/topic/Balto-Slavic-languages.

Tomaziefski LA (July 6 2020) Predicting Aptitude for Foreign Language Attainment. Independent study, Department of Data Science, Utica College.

Wong CH (2004) An analysis of factors predicting graduation of students at Defense Language Institute Foreign Language Center. Master's thesis, Operations Research Department, Naval Postgraduate School, Monterey, CA, http://hdl.handle.net/10945/1296.

Wordminds (2019) Nordic languages: what's the difference between all of them? Accessed January 20, 2021, https://wordminds.com/blog/difference-nordic-languages/.

Wright M, Ziegler A (2017) ranger: A fast implementation of random forests for high dimensional data in C++ and R. Journal of Statistical Software 77(1), 1–17, https://www.jstatsoft.org/article/view/v077i01.

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California